ON THE ROLE OF ATTENTION IN PROMPT-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Prompt-tuning is an emerging strategy to adapt large language models (LLM) to downstream tasks by learning a (soft-)prompt parameter from data. Despite its success in LLMs, there is limited theoretical understanding of the power of prompt-tuning and the role of the attention mechanism in prompting. In this work, we explore prompt-tuning for one-layer attention architectures and study contextual mixture-models where each input token belongs to a context-relevant or -irrelevant set. We isolate the role of prompt-tuning through a self-contained prompt-attention model. Our contributions are as follows: (1) We show that softmax-prompt-attention is provably more expressive than softmax-self-attention and linear-prompt-attention under our contextual data model. (2) We analyze the initial trajectory of gradient descent and show that it learns the prompt and prediction head with near-optimal sample complexity and demonstrate how prompt can provably attend to sparse context-relevant tokens. We also provide experiments that verify our theoretical insights on real datasets and demonstrate how prompt-tuning enables the model to attend to context-relevant information.

1 INTRODUCTION

Prompt-tuning provides a more efficient (cheaper/faster) alternative to fine-tuning the entire weights of the transformer by instead training (fewer) so-called prompt parameters that are appended on the input and can be thought of as an input interface. In fact, several recent works have demonstrated experimentally that prompt-tuning is not only more efficient, but often even becomes competitive to fine-tuning in terms of accuracy (Lester et al., 2021; Liu et al., 2023; Li and Liang, 2021). However, there is currently limited formal justification of such observations. This motivates our first question:

How does prompt-tuning compare to fine-tuning in terms of expressive power? Are there scenarios prompt-tuning outperforms fine-tuning in that regard?

The core constituent of a transformer, and thus of prompt-tuning, is the attention mechanism. Through the attention layer, prompts get to interact with other input features, create/modify attention weights, and facilitate the model to attend on latent task-specific information. The standard attention layer relies on softmax nonlinearities. Operationally, the softmax function allows a model to selectively focus on certain parts of the input tokens when generating attention outputs. However, there is little rigorous understanding of attention-based prompt-tuning. Concretely:

What is the role of the softmax-attention in prompt-tuning in terms of optimization and generalization? How does it locate and extract relevant contextual information?

Contributions. **(i)** We motivate a simplified *prompt* attention model naturally arising from self-attention with prompt tuning. We identify regimes where prompt-attention is more expressive than self-attention, demonstrating provable scenarios where prompt-tuning is superior to full-fine-tuning. We also rigorously show that prompt-attention is superior to linear-counterpars without softmax activation. **(ii)** By studying optimization and generalization dynamics of the initial trajectory of gradient descent for optimizing prompt attention, we find that a few iterations suffice to learn the prompt and prediction head with near-optimal sample complexity and high accuracy. **(iii)** The analysis provides insights into the role of softmax during optimization, by showing that it enables provably attending to sparse context-relevant tokens, while ignoring noisy/nuisance tokens. **(iv)** We also characterize the exact finite sample performance of prompt-attention assuming known prompt

but unknown prediction head. This reveals fundamental performance limits and the precise benefit of context information. **(v)** We uncover tradeoffs between model parameters such as the role of sparsity (i.e. fraction of context-relevant tokens), and the relative effects of the different constituents of context-relevant tokens. **(vi)** Finally, we empirically validate our findings on both synthetic contextual-mixture datasets and image-classification datasets. We compared multiple variants of prompt tuning against standard fine-tuning on the latter. Furthermore, we highlight the role of prompt-attention in selecting relevant tokens in the image classification setting.

2 Results

2.1 Motivation: Prompt-tuning

Consider a single-head self-attention layer $\boldsymbol{O}_{\text{pre}} = \phi(\boldsymbol{X}\boldsymbol{W}_Q\boldsymbol{W}_K^\top\boldsymbol{X}^\top)\boldsymbol{X}\boldsymbol{W}_V$ with input $\boldsymbol{X} \in \mathbb{R}^{T\times d}$ consisting of T tokens of dimension d each, trainable parameters $(\boldsymbol{W}_K, \boldsymbol{W}_Q, \boldsymbol{W}_V)$ and softmax nonlinearity. We scalarize the output of the self-attention output with a trainable linear head $\bar{\boldsymbol{U}}$ which yields $y_{\text{pre}} = \langle \bar{\boldsymbol{U}}, \boldsymbol{O}_{\text{pre}}\rangle = \langle \bar{\boldsymbol{U}}, \phi(\boldsymbol{X}\boldsymbol{W}_Q\boldsymbol{W}_K^\top\boldsymbol{X}^\top)\boldsymbol{X}\rangle$ Note here that we implicitly subsume the value matrix $\boldsymbol{W}_V$ in the linear head via $\boldsymbol{U} := \bar{\boldsymbol{U}}\boldsymbol{W}_V^\top$. We assume that the model above is pre-trained so that $\boldsymbol{W}_K, \boldsymbol{W}_Q, \boldsymbol{U}$ are fixed. Our goal is to use the pretrained transformer on (potentially) new classification tasks. Towards this goal, we explore the use of prompt-tuning, introduced in Li and Liang (2021); Lester et al. (2021) as an alternative to fine-tuning the existing transformer weights.

Prompt-tuning appends a trainable prompt $\boldsymbol{P} \in \mathbb{R}^{m\times d}$ to the input features $\boldsymbol{X} \in \mathbb{R}^{T\times d}$ with the goal of conditioning the transformer to solve the new classification task. Let $\boldsymbol{X}_{\boldsymbol{P}} := [\boldsymbol{P}^\top\ \boldsymbol{X}^\top]^\top \in \mathbb{R}^{(T+m)\times d}$ be the new transformer input. The output of the attention-layer is thus is of the form $\boldsymbol{O} = \phi(\boldsymbol{X}_{\boldsymbol{P}}\boldsymbol{W}_Q\boldsymbol{W}_K^\top\boldsymbol{X}^\top)\boldsymbol{X}$. Note that this is slightly different from $\boldsymbol{O}_{\text{pre}}$ in that now the layer computes a cross-attention between the augmented inputs $\boldsymbol{X}_{\boldsymbol{P}}$ and the original inputs $\boldsymbol{X}$. This is also equivalent to self-attention on $\boldsymbol{X}_{\boldsymbol{P}}$ after masking the prompt $\boldsymbol{P}$ from keys. This masking is used to cleanly isolate the residual contribution of the prompt without impacting the frozen attention output. Concretely, let $\boldsymbol{W}$ be the prediction head associated with the prompt tokens. As before, we scalarize the output by projecting with a linear head of size $(T+m)\times d$ as follows:

$$y = \langle [\boldsymbol{W}^\top\ \boldsymbol{U}^\top]^\top, \phi(\boldsymbol{X}_{\boldsymbol{P}}\boldsymbol{W}_Q\boldsymbol{W}_K^\top\boldsymbol{X}^\top)\boldsymbol{X}\rangle = \underbrace{\langle \boldsymbol{W}, \phi(\boldsymbol{P}\boldsymbol{W}_Q\boldsymbol{W}_K^\top\boldsymbol{X}^\top)\boldsymbol{X}\rangle}_{\text{prompt-attention } y_{\text{new}}} + \underbrace{\langle \boldsymbol{U}, \phi(\boldsymbol{X}\boldsymbol{W}_Q\boldsymbol{W}_K^\top\boldsymbol{X}^\top)\boldsymbol{X}\rangle}_{\text{frozen self-attention } y_{\text{pre}}}.$$

Here, y_{new} captures the additive impact of prompt-tuning on the prediction. We denote the trainable parameters in the model above as $\boldsymbol{\theta} := (\boldsymbol{W}, \boldsymbol{P})$. Since the y_{new} term becomes a self-contained module and the features attend directly to the prompt vector, we will refer to it as *prompt-attention*.

Our goal is understanding expressivity, training dynamics, and generalization of the above model. To simplify our analysis, we consider the following setting. (i) We focus our attention on the novel component y_{new} so as to isolate and fully understand the capabilities of prompt-attention. (ii) We assume $\boldsymbol{W}_K, \boldsymbol{W}_Q \in \mathbb{R}^{d\times d}$ are full-rank. (iii) We assume a single trainable prompt $\boldsymbol{q} \in \mathbb{R}^d$ i.e., $m = 1$.

Prompt-attention model. Using these assumptions and setting $\boldsymbol{q} := \boldsymbol{W}_K\boldsymbol{W}_Q^\top\boldsymbol{P}^\top \in \mathbb{R}^d$ and $\boldsymbol{w} = \boldsymbol{W}^\top \in \mathbb{R}^d$, we arrive at our core *prompt-attention* model

$$f_{\boldsymbol{\theta}}(\boldsymbol{X}) = \langle \boldsymbol{w}, \boldsymbol{X}^\top\phi(\boldsymbol{X}\boldsymbol{q})\rangle, \quad \boldsymbol{\theta} = (\boldsymbol{w}, \boldsymbol{q}). \tag{1}$$

We shall see that this model exhibits exciting properties to learn rich contextual relationships within the data and can even be more expressive than a single self-attention layer.

We remark that the model above is of interest even beyond the context of prompting: the prompt-attention model in (1) is reminiscent of the simplified model proposed in earlier seq2seq architectures Bahdanau et al. (2014); Xu et al. (2015); Chan et al. (2015) preceding self-attention and TFs Vaswani et al. (2017). Indeed, in the simplified attention mechanism of Bahdanau et al. (2014); Xu et al. (2015); Chan et al. (2015), the tokens' *relevance scores* and corresponding *attention weights* are determined by $\boldsymbol{a} = \phi(\boldsymbol{X}\boldsymbol{q})$ in which $\boldsymbol{q}$ is a trainable vector and ϕ is the softmax score transformation. Note here that the trainable parameter $\boldsymbol{q}$ corresponds exactly to the trainable prompt vector in (1).

2.2 CONTEXTUAL DATA MODEL

Dataset model. Consider supervised classification on IID data $(\boldsymbol{X}, y) \sim \mathcal{D}$ with features $\boldsymbol{X} \in \mathbb{R}^{T\times d}$ and binary label $y \in \{\pm 1\}$. The labels y are distributed as $\mathbb{P}(y=1) = 1-\mathbb{P}(y=-1) = \pi$; for simplicity, we set $\pi = 1/2$ so that $\mathbb{E}[y] = 0$. The tokens $\boldsymbol{x}_t, t \in [T]$ of input example $\boldsymbol{X} := [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T]$ are split into a *context-relevant* set $\mathcal{R} \subset [T]$ and *context-irrelevant* set $\mathcal{R}^c := [T] - \mathcal{R}$. Specifically, conditioned on the labels and relevance set $\mathcal{R}$, tokens $\boldsymbol{x}_t, t \in [T]$ within $\boldsymbol{X}$ are i.i.d. as follows

$$\boldsymbol{x}_t|y = \begin{cases} \boldsymbol{q}_\star + y\boldsymbol{w}_\star & , t \in \mathcal{R} \quad \text{(relevant token)} \\ -\delta^q \boldsymbol{q}_\star - \delta^w y \boldsymbol{w}_\star + \boldsymbol{z}_t & , t \notin \mathcal{R} \quad \text{(\textbf{ir}relevant token)}. \end{cases} \tag{DATA}$$

In the above, $\boldsymbol{q}_\star$ is a context-vector indicating token relevance and $\boldsymbol{w}_\star$ is a regressor vector. $y, \delta := (\delta^q, \delta^w), (\boldsymbol{z}_t)_{t=1}^T$ are independent variables as follows: • $\delta = (\delta^q, \delta^w) \in \mathbb{R}_+ \times \mathbb{R}$ is a bounded random variable that reflects *out-of-context* information within irrelevant tokens. However, δ is allowed to expose label information through δ^w. When $\delta = (0,0)$, we refer to (DATA) as **core dataset model**. • $\boldsymbol{z}_t$ are independent centered σ-subgaussian vectors with covariance $\boldsymbol{\Sigma}$, symmetric distribution and zero-third moment $\mathbb{E}\big[\boldsymbol{z}_T \otimes (\boldsymbol{z}_t^\top \boldsymbol{z}_t)\big] = 0$. When $\boldsymbol{\Sigma} = 0$, we refer to (DATA) **discrete dataset model**.

We allow the relevance set $\mathcal{R}$ to be non-stochastic. This includes $\mathcal{R}$ being adversarial to the classification model. We assume constant fraction $\zeta = |\mathcal{R}|/T \in (0,1)$ of label-relevant tokens for each input example $\boldsymbol{X}$ drawn from $\mathcal{D}$. Thus, ζ represents the sparsity of relevant signal tokens.

Training dataset $\mathcal{S} := (\boldsymbol{X}_i, y_i)_{i=1}^n$. We draw n i.i.d. samples from $\mathcal{D}$ to form our training dataset $\mathcal{S} := (\boldsymbol{X}_i, y_i)_{i=1}^n$. For i'th example $(y_i, \boldsymbol{X}_i)$, we denote the tokens by $(\boldsymbol{x}_{i,t})_{t=1}^T$, noise by $(\boldsymbol{z}_{i,t})_{t=1}^T$, relevance set by $\mathcal{R}_i$, and out-of-context variable by $\delta_i = (\delta_i^q, \delta_i^w)$. Ideally, for each i, we wish to identify its context-relevant set $\mathcal{R}_i$ and discard the rest: This would especially help when the SNR is small, i.e. $\zeta = |\mathcal{R}_i|/T \ll 1$. This is precisely the role of the context-vector $\boldsymbol{q}_\star$: Observe that relevant tokens have positive correlation with $\boldsymbol{q}_\star$ whereas irrelevant tokens have non-positive correlation with $\boldsymbol{q}_\star$ in expectation. Thus, by attending based on $\boldsymbol{q}_\star$ correlation, we can select relevant tokens.

2.3 POWER OF PROMPT-TUNING ON DISCRETE DATASET

Denote $W = \|\boldsymbol{w}_\star\|, Q = \|\boldsymbol{q}_\star\|$, $\bar{\boldsymbol{w}}_\star = \boldsymbol{w}_\star/W$, $\bar{\boldsymbol{q}}_\star = \boldsymbol{q}_\star/Q$. To proceed, we first study the discrete dataset where $\boldsymbol{\Sigma} = 0$ and correlation coefficient $\rho = \bar{\boldsymbol{w}}_\star^\top \bar{\boldsymbol{q}}_\star$ obeys $|\rho| < 1$. First note that if $\delta = (\delta^q, \delta^w)$ admits a single value, even a linear model $f^{\text{LIN}}(\boldsymbol{X}, \boldsymbol{w}) := \boldsymbol{w}^\top \boldsymbol{X} \mathbb{1}_T$ can solve the problem.

Observation 1 (Linear model)**.** *Suppose $(\delta^q, \delta^w) = (\Delta^q, \Delta^w)$ almost surely for $\Delta^w \neq \zeta/(1-\zeta)$. Set $\boldsymbol{w}'_\star = (\boldsymbol{I} - \bar{\boldsymbol{q}}_\star \bar{\boldsymbol{q}}_\star^\top)\boldsymbol{w}_\star$. Either $f^{\text{LIN}}(\boldsymbol{w}'_\star)$ or $f^{\text{LIN}}(-\boldsymbol{w}'_\star)$ achieves perfect accuracy.*

Thus, to investigate the expressivity of f^{ATT} and f^{SATT}, (δ^q, δ^w) would need to admit two or more values. Perhaps surprisingly, we prove that, as soon as, (δ^q, δ^w) comes from a binary distribution, then f^{SATT} can indeed provably fail in the regime $\delta^q \geq 0$ where prompt-attention thrives.

Theorem 1 (Separation of Population Accuracy)**.** *The following statements hold for discrete dataset:*

*1. **Prompt Attention:** Suppose $\delta^q \geq 0$ and $|\delta^w| \leq C$ almost surely. Set $\boldsymbol{q}'_\star = (\boldsymbol{I} - \bar{\boldsymbol{w}}_\star \bar{\boldsymbol{w}}_\star^\top)\boldsymbol{q}_\star, \boldsymbol{w}'_\star = (\boldsymbol{I} - \bar{\boldsymbol{q}}_\star \bar{\boldsymbol{q}}_\star^\top)\boldsymbol{w}_\star$. Choosing $\boldsymbol{\theta} = (\boldsymbol{w}'_\star, \Gamma \boldsymbol{q}'_\star)$, $f^{\text{ATT}}_{\boldsymbol{\theta}}$ achieves perfect accuracy for large $\Gamma > 0$.*

*2. **Self-attention:** In (DATA), choose (δ^q, δ^w) to be $(0,0)$ or (Δ, Δ) equally-likely with $\Delta > (1-\zeta)^{-2}$. For any choice of $(\boldsymbol{U}, \boldsymbol{W})$, $f^{\text{SATT}}(\boldsymbol{U}, \boldsymbol{W}) = \langle \boldsymbol{U}, \phi(\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^\top)\boldsymbol{X} \rangle$ achieves 50% accuracy.*

*3. **Linear Prompt Attention:** Choose (δ^q, δ^w) to be $(0,0)$ or $(\Delta, -\Delta)$ equally-likely with $\Delta > \sqrt{\zeta/(1-\zeta)}$. For any $(\boldsymbol{w}, \boldsymbol{q})$, $f^{\text{LIN-ATT}}(\boldsymbol{w}, \boldsymbol{q}) = \boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{q}$ achieves at most 75% accuracy.*

While surprising, the reason prompt-attention can provably beat self-attention is because it is optimized for context-retrieval and can *attend* perfectly on the relevant contextual information. In contrast, self-attention scores are fully feature-based; thus, context information is mixed with other features and can be lost during aggregation of the output.

2.4 GRADIENT-BASED ANALYSIS OF PROMPT ATTENTION

Our GD analysis concerns the prompt-attention model $f^{\text{ATT}}_{\boldsymbol{\theta}}$, so we simply write f_θ. Also, without any further explicit reference, we focus on the core dataset model, i.e. (DATA) with $\delta = (0,0)$. All our results hold under the mild correlation assumption $|\rho| < W/Q$ and for simplicity we assume

isotropic noise $\boldsymbol{\Sigma} = \sigma^2\mathbb{I}$ and handle the general case in the appendix. We use square-loss $\widehat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\theta}) := \frac{1}{2n}\sum_{i=1}^{n}(y_i - f_{\boldsymbol{\theta}}(\boldsymbol{X}_i))^2$. For data generated from (DATA), we show that a three-step gradient-based algorithm achieves low **classification error**: $\mathrm{ERR}(\hat{\boldsymbol{\theta}}) := \mathbb{P}(yf_{\hat{\boldsymbol{\theta}}}(\boldsymbol{X}) > 0)$.

Algorithm: We split the train set in three separate subsets $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$ of size n each. Starting from $\boldsymbol{w}_0 = 0, \boldsymbol{q}_0 = 0$, the algorithm proceeds in three gradient steps for step sizes $\eta > 0$ and $\gamma > 0$ and a final debiasing step as follows: **(1)** $\widehat{\boldsymbol{w}}_1 := -\eta\nabla_{\boldsymbol{w}}\widehat{\mathcal{L}}_{\mathcal{S}_1}(0,0)$, **(2)** $\widehat{\boldsymbol{q}}_1 := -\gamma\nabla_{\boldsymbol{q}}\widehat{\mathcal{L}}_{\mathcal{S}_2}(0,\widehat{\boldsymbol{w}}_1)$, **(3)** $\widehat{\boldsymbol{w}}_2 := -\eta\nabla_{\boldsymbol{w}}\widehat{\mathcal{L}}_{\mathcal{S}_3}(\widehat{\boldsymbol{q}}_1,\widehat{\boldsymbol{w}}_1)$, where $\widehat{\mathcal{L}}_{\mathcal{S}_j}, j = 1,2,3$ is the loss evaluated on sets $\mathcal{S}_j$.

To gain intuition consider first the population counterpart of the algorithm, i.e., (3) substituting $\widehat{\mathcal{L}}(\boldsymbol{w},\boldsymbol{q})$ with its population version $\mathbb{E}\left[\widehat{\mathcal{L}}(\boldsymbol{w},\boldsymbol{q})\right]$ in all three steps. To see how the second population update $\boldsymbol{q}_1$ selects good tokens, we investigate the (normalized) relevance scores $\boldsymbol{x}_t^\top\boldsymbol{q}_1$ of signal vs noise tokens: Attending to context-relevant tokens requires the signal/context relevance scores be larger than the noisy ones. Concretely, suppose we have

$$B := \min_{t\in\mathcal{R}}(\boldsymbol{q}_\star + y\boldsymbol{w}_\star^\top)\boldsymbol{q}_1 \geq 2\max_{t\in\mathcal{R}^c}\boldsymbol{z}_t^\top\boldsymbol{q}_1. \tag{2}$$

Then, a large enough second gradient step (large γ) finds $\boldsymbol{q}_1$ that attends (nearly) perfectly to context-relevant tokens in $\mathcal{R}$ and attenuates (almost) all the noise tokens in $\mathcal{R}^c$:

$$a_t = [\phi(\boldsymbol{X}\boldsymbol{q}_1)]_t = e^{\gamma r_t}/S\begin{cases} = e^{\gamma B}/S \to 1/(\zeta T) & t\in\mathcal{R} \\ \leq e^{\gamma B/2}/S \to 0 & t\in\mathcal{R}^c \end{cases}.$$

Theorem 2 (Population)**.** *Consider the three-step algorithm with population gradients and assume $\boldsymbol{q}_\star \perp \boldsymbol{w}_\star$ for simplicity. For step-size η small enough, there exist sufficiently large $\alpha > 1$ and step-size γ such that* $\mathrm{ERR}\left(f_{\boldsymbol{\theta}^\gamma}\right) \leq 2/T^{\alpha-1}$ *, provided* $Q := \|\boldsymbol{q}_\star\| \geq \sigma\sqrt{8\log\left(2(1-\zeta)T^\alpha\right)}$.

As per our discussion, the theorem's condition for Q to be large guarantees (2) so that attention weights $\phi(\boldsymbol{X}\boldsymbol{q}_1)$ provably select the relevant tokens and discard those irrelevant.

Below we provide a finite-sample counterpart of Theorem 2. (This is a simplified version; see Section E for details). For clarity, we fix $\sigma \propto 1$ and use $\gtrsim, \tilde{\mathcal{O}}()$ to suppress logarithmic dependencies on T, n.

Theorem 3 (Finite-sample)**.** *Consider the three-step algorithm with an additional de-biasing step explained in Sec. E.3. Choose step sizes $\eta \propto 1/(Q^2\zeta)$ and large enough γ. Suppose $Q \gtrsim 1 + W$, $\zeta \leq 0.9$. Declare $\mathtt{rate}_{\mathrm{LIN}} = \zeta^2W^2T$ and error rate $\mathtt{rate} := Q^2 \wedge Q\sqrt{d} \wedge n^{1/3}\zeta^{2/3}(W/Q)^{4/3} \wedge (n/d)\mathtt{rate}_{\mathrm{LIN}}$. With probability $1 - ce^{-\tilde{\mathcal{O}}(\mathtt{rate}\wedge d)}$ over the training process,* $\mathrm{ERR}(f_{(\boldsymbol{\theta})}) \leq ce^{-\tilde{\mathcal{O}}(\mathtt{rate})}$.

$\mathtt{rate}_{\mathrm{LIN}}$ corresponds to the error rate of the linear baseline f^{LIN} (see Fact D.1). This appears in our theorem with $(n/d)\mathtt{rate}_{\mathrm{LIN}}$; thus, as soon as $n \propto d$, GD can beat the linear model. Remaining components of $\mathtt{rate}$ are regularity conditions: Analogous to its population-counterpart discussed above, we ask for Q to be large enough. We also need $n \gtrsim (W/Q)^4\zeta^2$ to suppress the variance due to the $\boldsymbol{q}_\star$ terms during perturbation analysis of $\hat{\boldsymbol{w}}_1$. By letting $n \to \infty$ in Theorem 3, we end up with error rate of $e^{-cQ(Q\wedge\sqrt{d})}$ which provably beats the rate of f^{LIN} whenever $Q(Q \wedge \sqrt{d}) \geq \zeta^2W^2T$.

Sharp error rates: Finally, in Appendix D we provide an exact analysis of the classification error when $\boldsymbol{q}_\star$ is known and only $\boldsymbol{w}_\star$ is estimated from data. This analysis exactly quantifies the value of context-information and how prompt-tuning retrieves it. Specifically, we prove a sharp asymptotic error rate of $\mathcal{Q}\left(\frac{e^{Q^2/4}}{\sqrt{1+\mathtt{ISNR}(n/d)}}\cdot\sqrt{\frac{\zeta^2W^2T}{1-\zeta}}\right)$ where $\mathtt{ISNR}(\alpha) := \alpha^{-1}\frac{(1-\zeta)e^{-Q^2/2}}{\mathtt{rate}_{\mathrm{LIN}}}$, $\mathcal{Q}(\cdot)$ is the gaussian tail function and noise is isotropic gaussian. This strictly improves over the existing optimal rates for (context-free) Gaussian mixture models thanks to the context information.

DISCUSSION

We initiate a theoretical investigation of prompt-tuning. Our results suggest many interesting future directions including (1) extension to deeper architectures by characterizing the role of softmax-attention in individual layers, (2) developing a stronger theoretical and empirical understanding of when/if prompt-tuning is superior to fine-tuning, (3) extending our model to include multiple prompt vectors (and perhaps extending (DATA) to include multiple context vectors), and (4) investigating the role of multiple attention heads in prompt-tuning.

REFERENCES

A theoretical understanding of vision transformers.

https://openai.com/blog/chatgpt/.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Pierre Baldi and Roman Vershynin. The quarks of attention. *arXiv preprint arXiv:2202.08371*, 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

William Chan, Navdeep Jaitly, Quoc V Le, and Oriol Vinyals. Listen, attend and spell. *arXiv preprint arXiv:1508.01211*, 2015.

Mostafa Dehghani, Alexey Gritsenko, Anurag Arnab, Matthias Minderer, and Yi Tay. Scenic: A jax library for computer vision research and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21393–21398, 2022.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning*, pages 2793–2803. PMLR, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. *arXiv preprint arXiv:2110.10090*, 2021.

Tolga Ergen, Behnam Neyshabur, and Harsh Mehta. Convexifying transformers: Improving optimization and understanding of transformer networks. *arXiv preprint arXiv:2211.11052*, 2022.

Samy Jelassi, Michael Eli Sander, and Yuanzhi Li. Vision transformers provably learn spatial structure. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=eMW9AkXaREI.

Stefani Karp, Ezra Winston, Yuanzhi Li, and Aarti Singh. Local signal adaptivity: Provable feature learning in neural networks beyond kernels. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 24883–24897. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/d064bf1ad039ff366564f352226e7640-Paper.pdf.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer Science & Business Media, 1991.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

Hesameddin Mohammadi, Armin Zare, Mahdi Soltanolkotabi, and Mihailo R Jovanović. Convergence and sample complexity of gradient methods for the model-free linear quadratic regulator problem. *arXiv preprint arXiv:1912.11899*, 2019.

Deepak Narayanan, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Patwary, Vijay Korthikanti, Dmitri Vainbrand, Prethvi Kashinkunti, Julie Bernauer, Bryan Catanzaro, et al. Efficient large-scale language model training on gpu clusters using megatron-lm. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–15, 2021.

David Pollard. Empirical processes: theory and applications. Ims, 1990.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

Arda Sahiner, Tolga Ergen, Batu Ozturkler, John Pauly, Morteza Mardani, and Mert Pilanci. Unraveling attention via convex duality: Analysis and interpretations of vision transformers. *International Conference on Machine Learning*, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ByxRM0Ntvr.

A Related Works

Attention, specifically the so-called Self-attention, is the backbone mechanism of transformers (Vaswani et al., 2017). It differs from conventional models (e.g., multi-layer perceptrons (MLPs) and convolutions neural networks (CNNs)) in that it computes feature representations by globally modeling interactions between different parts of an input sequence. Despite tremendous empirical success (see, e.g., Vaswani et al., 2017; Brown et al., 2020; Saharia et al., 2022; Ramesh et al., 2022; cha; Narayanan et al., 2021; Reed et al., 2022, and references therein), the underlying mechanisms of the attention layer remain largely unknown: How does it learn? What makes it better (and when) compared to conventional architectures? Yun et al. (2020) show that self-attention based transformers

with large enough depth are universal approximators of seq2seq functions. Focusing on the self-attention component, Edelman et al. (2021) show that self-attention can efficiently represent sparse functions of its input space, while Sahiner et al. (2022); Ergen et al. (2022) analyze convex-relaxations of Self-attention, and Baldi and Vershynin (2022); Dong et al. (2021) study the expressive ability of attention layers. However, these works do *not* characterize the optimization and generalization dynamics of attention. To the best of our knowledge, the only prior works attempting this are Jelassi et al. (2022) and ICL. Jelassi et al. (2022) assume a simplified attention structure in which the attention matrix is *not* directly parameterized in terms of the input sequence. An interesting recent ICLR submission (ICL) studies optimization/generalization of training a full self-attention model. Our paper differs in a variety of ways including: (1) the data model in ICL is different as it has no notion of context vector and the noise in the data is assumed to be bounded whereas our model captures the role of context and our noise model is sub-Gaussian. (2) unlike ICL, we develop a precise asymptotic analysis that reveals the precise role of various problem parameters. (3) Finally, our focus here is on understanding prompt tuning via prompt-attention and when it can potentially improve upon vanilla self-attention (which is the focus of ICL).

B Experiments

First, we verify the utility of the prompt attention via experiments on a synthetic setting that precisely follows the contextual data model (cf. Section 2.2). Subsequently, we explore prompt-tuning on various image-classification tasks that are motivated by the contextual data model and compare it with the standard fine-tuning method. Finally, we validate the utility of prompt vectors in distinguishing relevant tokens from irrelevant tokens via prompt attention under an image classification setting.

B.1 Synthetic experiments

Here, we generate a synthetic dataset according to the *core dataset model*, i.e., we have $\delta = (\delta^q, \delta^w) = (0, 0)$ for all examples in the dataset. In particular, we consider a setting with $T = 500$, d=50, and $\zeta = 0.1$, i.e., each example has 500 tokens out of which 10% tokens are relevant. As for the noisy tokens, they consists of i.i.d. $\mathcal{N}(0, \mathbf{I})$ vectors. Assuming that $\boldsymbol{q}_\star \perp \boldsymbol{w}_\star$ and $\sqrt{T}W = 3$, we generate $n = 10 \cdot d$ training examples from the core dataset model for varying Q. Fig. 1 showcases the performance of prompt attention when combined with the estimates $\hat{\boldsymbol{q}}$ and $\hat{\boldsymbol{w}}$ gradient-based algorithm in Section E. We also showcase the performance of the linear model and two oracle methods where we assume access to true $\boldsymbol{q}_\star$ and true $(\boldsymbol{q}_\star, \boldsymbol{w}_\star)$ while applying the prompt attention, respectively.

Note that prompt attention achieves a vanishing classification test error in this setting whereas a natural baseline (linear model) can fail to achieve a good performance. On the other hand, the prompt attention enable by the gradient-based algorithm successfully achieves a high accuracy as the context energy (defined by Q) increases, validating the utility of prompt attention as well as our gradient-based algorithm.

B.2 Image classification experiments

B.2.1 Experimental setup

Dataset. Motivated by our contextual data model (cf. Section 2.2), we construct multiple datasets based on the CIFAR-10 dataset (Krizhevsky et al., 2009) to conduct our evaluation. We construct three such datasets (see Fig. 2 for examples).

- Full-Tiled. For this dataset, each examples consists of a 64x64 images obtained by arranging a 32x32 image from the original CIFAR-10 dataset a tiling pattern with four tiles (cf. Fig. 2a).
- Partial-Tiled. This is dataset is similar to Full-Tiled with the exception that each image has at-least T out of 4 tiles replaced by a patch of i.i.d. random Gaussian noise with mean zero and variance 0.2. Note that $T \in \{1, 2, 3\}$ is a random number as well as the location of the noisy tiles for each image in the dataset (cf. Fig. 2b).
- Embed-in-ImageNet. In this dataset (Karp et al., 2021), we construct an example by simply embedding a 32 x 32 image from CIFAR-10 at a random location in a 64 x 64 background corresponding to a randomly selected image from ImageNet Russakovsky et al. (2015). We

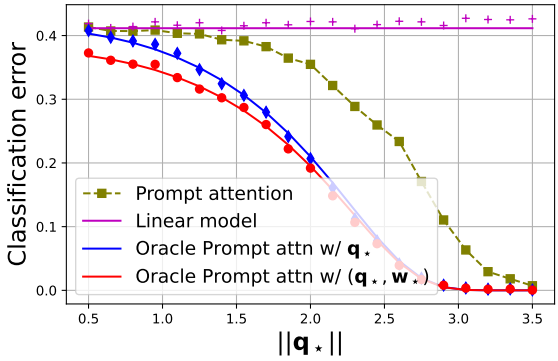
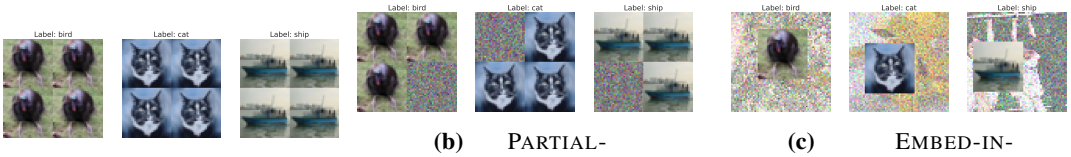
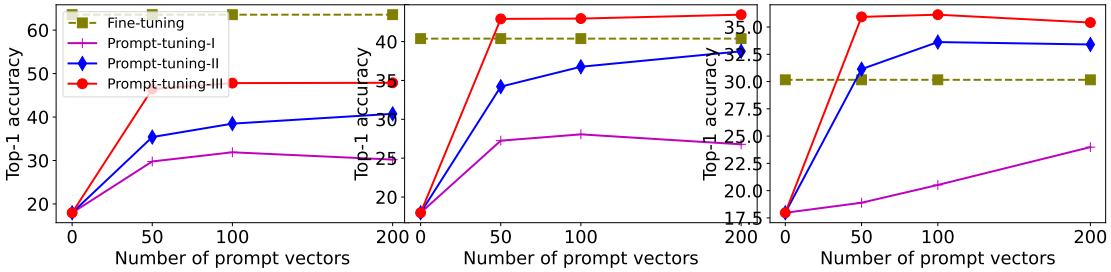


Figure 1: Performance of prompt attention on the synthetic setting described in Section B.1. For prompt attention, we employ the gradient-based algorithm from Section E to obtain estimates $\hat{q}$ and $\hat{w}$. For the baseline linear model and two oracle settings, we have closed-form expressions for their asymptotic test error (cf. Theorem 4), which is depicted by solid lines. On the other hand, markers show the finite sample performance of these three methods. All finite sample performances are obtained by averaging over 20 independent runs.



(a) FULL-TILED **(b)** PARTIAL-TILED **(c)** EMBED-IN-IMAGENET

Figure 2: Illustration of different CIFAR-10 based datasets utilized in image classification experiments (cf. Section B.2). Note that all three variants correspond to 10-way multiclass classification tasks corresponding to 10 original classes in CIFAR-10.

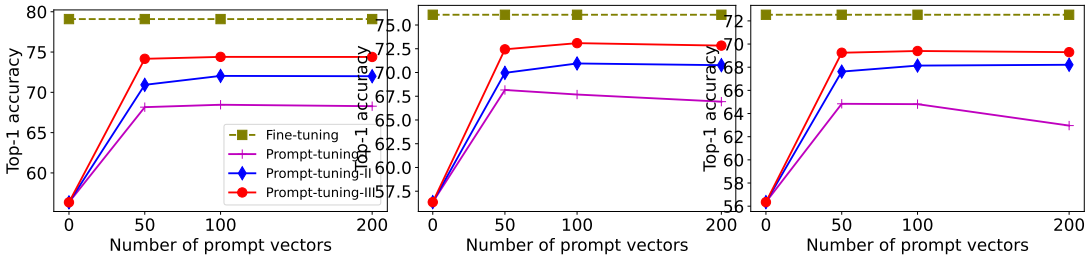


(a) Full dataset **(b)** Capped 10% **(c)** Capped 2%

Figure 3: Performance of fine-tuning vs. prompt-tuning on 10-way classification tasks defined by EMBED-IN-IMAGENET dataset. Full dataset has 50K training examples. Capped 10% and 2% correspond to sub-sampled *train* sets where we select exactly 500 and 100 examples per class from the full dataset. Note that number of prompt vectors equal to 0 corresponds to *zero-shot* performance.

also add a i.i.d. random Gaussian noise with mean zero and variance 0.2 to the background (cf. Fig. 2c).

By construction, each dataset has 50,000 train and 10,000 test examples corresponding to train and test set of CIFAR-10. We also considered data-limited setting where we keep the test set intact but reduce the size of train set by sampling a fixed number of images for each class from the original train set. Note that all three datasets define 10-way multiclass classification tasks with CIFAR-10 classes as potential labels.



(a) Full dataset **(b)** Capped 10% **(c)** Capped 2%

Figure 4: Performance of fine-tuning vs. prompt-tuning on 10-way classification tasks defined by PARTIAL-TILED dataset. Full dataset has 50K training examples. Capped 10% and 2% correspond to sub-sampled *train* sets where we select exactly 500 and 100 examples per class from the full dataset. Note that number of prompt vectors equal to 0 corresponds to *zero-shot* performance.

Model architecture. We utilize a tiny variant of the Vision transformer model Dosovitskiy et al. (2021) for all of our experiments in this subsection. This variant has 12 transformer layers with it hidden dimension, MLP intermediate dimension, and number of heads per attention layer being equal to 192, 768, and 3, respectively. The patch size in our study is set to be 4x4. The model itself (without counting the trainable parameters/weights during prompt-tuning) has approximately 5.44M parameters. We rely on the CLS token to obtain the classification logits.

Training. We rely on Scenic library (Dehghani et al., 2022)[1] to conduct our experiments on image classification. Following the default settings in the library along with a coarse grid search, we employ Adam optimizer (Kingma and Ba, 2014) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay = 0.1, and batch size = 128 while training a *randomly initialized* model. Furthermore, we employ a linear warm-up of learning rate followed cosine learning rate schedule with base learning rate 3e-3. As for the fine-tuning and prompt-tuning experiments that (partially) initialize from a *pre-trained* model, we rely on SGD with momentum parameter 0.9 and batch size = 128 to update trainable parameters. Again, we utilize a linear warm-up of learning rate followed cosine learning rate schedule. Throughout our experiments, the base learning rates for fine-tuning and prompt-tuning are 1e-3 and 0.1, respectively.

Methods. In our fine-tuning experiments, we update all pre-trained model parameters. As for prompt-tuning, we only update newly introduced (prompt) variables and keep the pre-trained network frozen. We consider three variants of prompt-tuning: 1) PROMPT-TUNING-I (Lester et al., 2021), where we add trainable vector between CLS token embedding and first image (patch) embeddings only at the input; 2) PROMPT-TUNING-II (Li and Liang, 2021), where we add the *same* trainable vectors between the CLS embedding and the first image embedding at the input of every transformer layer; and 3) PROMPT-TUNING-III, where we add *different* trainable vectors between the CLS embedding and the first image patch embedding at the input of every transformer layer. Note that the number of trainable parameters in PROMPT-TUNING-I and PROMPT-TUNING-II do not scale with the number of layers whereas we have linear scaling with number of layers in PROMPT-TUNING-III. Interestingly, all three prompt-tuning variant are identical when the number of layers is 1, which corresponds to the setup we theoretically analyzed in the paper. However, they exhibit remarkably different behavior for a multi-layer transformer model, as we show in the next subsection.

B.2.2 RESULTS

Here, the main goal of our exploration is to highlight the different behavior of fine-tuning and prompt-tuning. Towards this, we utilize a model trained on FULL-TILED dataset as the pre-trained model. This model achieves top-1 (in-domain) accuracy of 80.43 on FULL-TILED test set. In contrast, it achieves *zero-shot* top-1 accuracy of 56.35 and 17.97 on PARTIAL-TILED and EMBED-IN-IMAGENET,

[1] https://github.com/google-research/scenic

respectively. This alludes to the fact that EMBED-IN-IMAGENET corresponds to a larger distribution shift from the pre-training distribution (FULL-TILED), as compared to PARTIAL-TILED.

Fig. 3 and Fig. 4 showcase the performance of fine-tuning and prompt-tuning approaches on EMBED-IN-IMAGENET and PARTIAL-TILED, respectively. Note that fine-tuning outperforms prompt-tuning in a data-rich setting (cf. Fig. 3a and 4a). This is due to fine-tuning having the ability to update a large number of model parameters (5.4M in our case). In contrast, with 2000 prompt vectors, PROMPT-TUNING-III (the most expensive prompt-tuning method out of all three) only updates 460.8K parameters.

Interestingly, in the data-limited regimes, prompt-tuning becomes more competitive with fine-tuning. In fact, the best performing prompt-tuning method outperforms the fine-tuning (cf. Fig. 3b and 3c) on EMBED-IN-IMAGENET, where fine-tuning can easily overfit as it cannot leverage the benefits of the pre-training data due to a large distribution-shift between FULL-TILED and EMBED-IN-IMAGENET.

As for the varying performance among different prompt-tuning approaches, part of the performance gap between PROMPT-TUNING-III and PROMPT-TUNING-II can be attributed to the larger number of trainable parameters available to PROMPT-TUNING-III. Even more interestingly, PROMPT-TUNING-II consistently outperforms PROMPT-TUNING-I, even with the same number of trainable parameters. This alludes to the fact that optimization and architecture choices play a major role beyond just the number of trainable parameters. As mentioned earlier, our theoretical treatment for a single-layer model cannot distinguish among these different prompt-tuning approaches. As a result, we believe that our empirical observations point to multiple interesting avenues for future work.

B.3 ATTENTION WEIGHTS FOR PROMPT VECTORS

Finally, we explore what role prompt-attention, i.e., the attention weights with prompt vectors as keys and image patches/tokens as values, plays towards underlying task. In Fig. 5, we illustrate one representative example. The figures shows how average attention weights from prompt vectors to image tokens/patches evolve across transformer layers, when we employ PROMPT-TUNING-III. Indeed, the figure verifies that prompt-attention helps locate the relevant tokens/patches from the irrelevant patches, validating our starting hypothesis in Section 2.1 and 2.2.

C FORMAL STATEMENT OF ASSUMPTIONS AND NOTATIONS

First, we formally state our assumptions on the noisy tokens. The more general condition is that noise is subgaussian and satisfies a mild zero third-moment condition.

Assumption 1.a. *The noise vector $\boldsymbol{z} \sim \mathcal{SN}(\sigma)$ is centered σ-subGaussian, i.e. $\|\boldsymbol{z}\|_{\psi_2} = \sigma$. Moreover, its distribution is symmetric and zero-third moment, i.e. $\mathbb{E}\big[\boldsymbol{z} \otimes (\boldsymbol{z}^\top \boldsymbol{z})\big] = 0$. We denote the noise covariance $\boldsymbol{\Sigma} := \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top]$.*

For some of our results it will be convenient to further assume that noise is Gaussian since this allows leads to precise formulas that are easily interpretable.

Assumption 1.b. *The noise vector $\boldsymbol{z} \sim \mathcal{N}(0, \sigma^2 \mathbb{I})$ is isotropic Gaussian with variance σ^2.*

Second, a mild assumption on the correlation between the context $\boldsymbol{q}_\star$ and classifier $\boldsymbol{w}_\star$, which guarantees pure signal tokens $\boldsymbol{q}_\star + y\boldsymbol{w}_\star$ are correctly classifier by the true regressor $\boldsymbol{w}_\star$, i.e. $y\boldsymbol{w}_\star^\top \boldsymbol{q}_\star + y\boldsymbol{w}_\star > 0$. For convenience, we denote $W := \|\boldsymbol{w}_\star\|, Q := \|\boldsymbol{q}_\star\|, \rho := \boldsymbol{q}_\star^\top \boldsymbol{w}_\star / \|\boldsymbol{q}_\star\| \|\boldsymbol{w}_\star\|$.

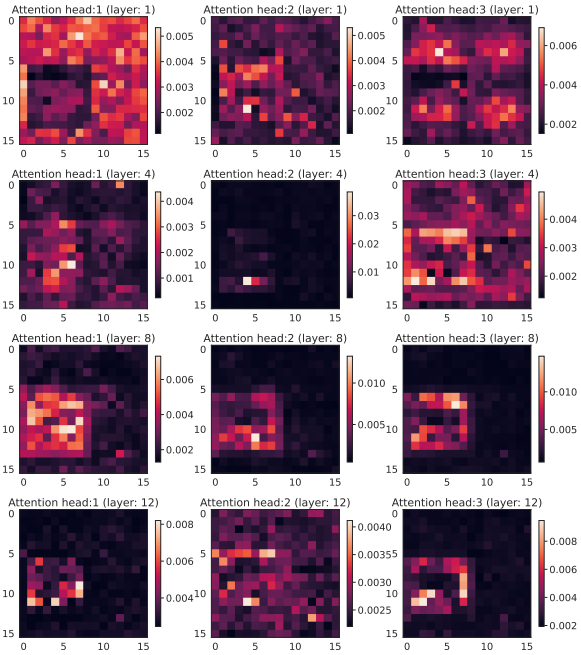
Assumption 3.a. *Correlation satisfies $|\rho| < \frac{W}{Q}$.*

We will also often consider the special case of zero correlation ρ and thus state it as separate assumption below. This orthogonality assumption, is useful for more tractable analysis as it helps decouple feature selection and prediction.

Assumption 3.b. *The context and classifier vectors are orthogonal, i.e. $\boldsymbol{q}_\star \perp \boldsymbol{w}_\star$.*

Notation. We use boldface letters for vectors and matrices. We use $\mathbb{1}_m$ to denote an m-dimensional vector of all ones. For a vector $\boldsymbol{v}$, $\|\boldsymbol{v}\|$ denotes its Euclidean norm and $\boldsymbol{v}/\|\boldsymbol{v}\|$ its normalization. $\phi(\cdot)$ denotes the softmax transformation. $Q(\cdot)$ denotes the tail function of standard normal distribution.

(a) Input image



(b) Average attention weights from prompts (keys) to image patches (values).

Figure 5: Illustration of how attention weights progressive change from the first layer (Figure 5b-top) to the last layer (Figure 5b-bottom) in the transformer model for a given input image (Figure 5a) when we employ PROMPT-TUNING-III. We plot average attention weights from 50 prompt vectors (keys) to 256 image patches (values). The attention weights for each attention head are naturally arranged in a 16 x 16 grid corresponding to the original locations of the patches in the image. Note that the attention weights in the early layer have a tiling pattern similar to that in FULL-TILED– the dataset utilized by the pre-trained model. However, as we progress deeper into the transformer, attention weights begin to capture the relevant patch locations in the dataset of interest, i.e., EMBED-IN-IMAGENET.

$\wedge$ / $\vee$ denotes the minimum / maximum of two numbers. $\tilde{\mathcal{O}}()$ and $\gtrsim$ notation suppress logarithmic dependencies. $\propto$ denotes proportionality.

D SHARP CHARACTERIZATION OF ACCURACY UNDER KNOWN CONTEXT

While the discrete dataset model is insightful, incorporating noise is crucial for understanding the fundamental limits of the benefits of context in attention. To this end, let us focus on the core dataset model where we set $\delta^q = \delta^w = 0$ and explore the role of noise in population accuracy. Also assume that noise is standard normal, i.e. Assumption 1.b.

• *Linear model.* The linear model aggregates tokens to obtain a simple Gaussian mixture distribution. Specifically, aggregated tokens are exactly distributed as $\frac{1}{T}\boldsymbol{X}^\top \mathbb{1}_T \sim \mathcal{N}(\zeta \boldsymbol{w}_\star, \frac{1-\zeta}{T}\boldsymbol{I})$, leading to the following well-known result.

Fact D.1. *For linear models, optimal accuracy obeys* $\min_{\boldsymbol{w}} \mathrm{ERR}(f^{\mathrm{LIN}}(\boldsymbol{w})) = Q(\sqrt{\frac{\zeta^2 W^2 T}{1-\zeta}})$ *where* $Q(\cdot)$ *is the tail function of the standard normal distribution.*

• *Prompt Attention model.* Since prompt-attention strictly generalizes the linear model, its accuracy is at least as good. The theorem below quantifies this and demonstrates that how context vector can enable an optimal weighting of relevant and irrelevant tokens to maximize accuracy. A general version of this theorem is proven under a non-asymptotic setting (finite T, d) as Theorem 9.

Theorem 4. *Consider the prompt-attention model* $f^{\mathrm{ATT}}_{\boldsymbol{\theta}}$. *Suppose* $\boldsymbol{w}_\star \perp \boldsymbol{q}_\star$ *and let* $\tau, \bar{\tau} > 0$ *be hyperparameters. Consider the following algorithm which uses the hindsight knowledge of* $\boldsymbol{q}_\star$ *to*

estimate $\boldsymbol{w}_\star$ and make prediction: Set $\hat{\boldsymbol{w}} = (\boldsymbol{I} - \bar{\boldsymbol{q}}_\star\bar{\boldsymbol{q}}_\star)\nabla\mathcal{L}_{\boldsymbol{w}}(0, \tau\bar{\boldsymbol{q}}_\star)$ and $\boldsymbol{\theta} = (\hat{\boldsymbol{w}}, \bar{\tau}\bar{\boldsymbol{q}}_\star)$. Suppose $\zeta^2W^2T, 1-\zeta, \alpha := n/d, e^{Q^2}, e^{\tau}$ each lie between two positive absolute constants. Suppose T is polynomially large in n and these constants and $\tilde{\mathcal{O}}(\cdot)$ hides polynomial terms in n. Define inverse-signal-to-noise-ratio*: $\texttt{ISNR}(\alpha, \tau) = \frac{(1-\zeta)e^{2\tau(\tau-Q)}}{\alpha\zeta^2W^2T}$. In the limit $T, d \to \infty$, the test error converges in probability to $\mathcal{Q}\left(\frac{e^{Q\bar{\tau}-\bar{\tau}^2}}{\sqrt{1+\texttt{ISNR}(\alpha,\tau)}} \cdot \sqrt{\frac{\zeta^2W^2T}{1-\zeta}}\right)$. In this limit, optimal hyperparameters are $\tau = \bar{\tau} = Q/2$ and leads to optimal $\texttt{ISNR}(\alpha) := \frac{(1-\zeta)e^{-Q^2/2}}{\alpha\zeta^2W^2T}$ and the error*

$$\text{ERR}(\alpha, Q, W) = \mathcal{Q}\left(\frac{e^{Q^2/4}}{\sqrt{1+\texttt{ISNR}(\alpha)}} \cdot \sqrt{\frac{\zeta^2W^2T}{1-\zeta}}\right)$$

Here, a few remarks are in place. Note that $\texttt{rate}_{\text{LIN}} := \sqrt{\frac{\zeta^2W^2T}{1-\zeta}}$ term is the population error rate of f^{LIN} from Fact D.1. In the limit $\alpha = n/d \to \infty$, the rate of f^{ATT} is cleanly given by $e^{Q/4}\texttt{rate}_{\text{LIN}}$ demonstrating the strict superiority of prompt-attention. Moreover setting $Q = 0$ (no prompt info), since feature-output of f^{LIN} (i.e. $\boldsymbol{X}^\top\phi(\boldsymbol{X}\mathbb{1})$) is (essentially) a binary Gaussian mixture distribution, our error rate recovers the Bayes-optimal f^{LIN} classifier which has a finite-sample rate of $\texttt{rate}_{\text{LIN}}/\sqrt{1+(1-\zeta)/(\alpha\zeta^2W^2T^2)}$. Prompt-tuning also strictly improves this because our $\texttt{ISNR}(\alpha)$ introduces an additional $e^{-Q^2/2}$ multiplier.

E Gradient-based analysis of prompt attention

This section investigates how gradient-descent optimization of the prompt-attention model learns (DATA). Concretely, it shows that a few gradient steps can provably attend on the context-relevant tokens leading to high-classification accuracy. Our results capture requirements on sample complexity in terms of all problem parameters, i.e. dimension d, correlation ρ, context / signal energies Q / W, number T of tokens, and sparsity ζ. This allows studying tradeoffs in different regimes.

Our analysis in this section concerns the prompt-attention model $f^{\text{ATT}}_{\boldsymbol{\theta}}$, so we simply write f_θ. Also, without any further explicit reference, we focus on the core dataset model, i.e. (DATA) with $\delta = (0, 0)$. All our results here hold under the mild noise and correlation assumptions: Assumption 1.a and Assumption 3.a. We will not further state these. Finally, for simplicity of presentation we assume here isotropic noise $\boldsymbol{\Sigma} = \sigma^2\mathbb{I}$ and handle the general case in the appendix.

E.1 Gradient-based algorithm

For data generated from (DATA), we show the three-step gradient-based algorithm described below achieves high test accuracy. Our analysis also explains why three appropriately chosen steps suffice.

Algorithm: We split the train set in three separate subsets $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$ of size n each. Starting from $\boldsymbol{w}_0 = 0, \boldsymbol{q}_0 = 0$, the algorithm proceeds in three gradient steps for step sizes $\eta > 0$ and $\gamma > 0$ and a final debiasing step as follows:

$$\widehat{\boldsymbol{w}}_1 := -\eta\nabla_{\boldsymbol{w}}\widehat{\mathcal{L}}_{\mathcal{S}_1}(0, 0), \tag{3a}$$

$$\widehat{\boldsymbol{q}}_1 := -\gamma\nabla_{\boldsymbol{q}}\widehat{\mathcal{L}}_{\mathcal{S}_2}(0, \widehat{\boldsymbol{w}}_1), \tag{3b}$$

$$\widehat{\boldsymbol{w}}_2 := -\eta\nabla_{\boldsymbol{w}}\widehat{\mathcal{L}}_{\mathcal{S}_3}(\widehat{\boldsymbol{q}}_1, \widehat{\boldsymbol{w}}_1), \tag{3c}$$

where $\widehat{\mathcal{L}}_{\mathcal{S}_j}, j = 1, 2, 3$ is the square-loss evaluated on sets $\mathcal{S}_j$.

E.2 Population analysis

To gain intuition we first present results on the population counterpart of the algorithm, i.e., (3) substituting $\widehat{\mathcal{L}}(\boldsymbol{w}, \boldsymbol{q})$ with its population version $\mathcal{L}(\boldsymbol{w}, \boldsymbol{q}) = \mathbb{E}\big[\widehat{\mathcal{L}}(\boldsymbol{w}, \boldsymbol{q})\big]$ in all three steps. It is convenient to introduce the following shorthand notation for the negative gradient steps $\boldsymbol{G}_{\boldsymbol{q}}(\boldsymbol{q}, \boldsymbol{w}) := -\nabla_{\boldsymbol{q}}\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = \mathbb{E}_{y,\boldsymbol{X}}\big[(y - f_{\boldsymbol{\theta}}(\boldsymbol{X}))\nabla_{\boldsymbol{q}}f_{\boldsymbol{\theta}}(\boldsymbol{X})\big]$ and $\boldsymbol{G}_{\boldsymbol{w}}(\boldsymbol{q}, \boldsymbol{w}) := -\nabla_{\boldsymbol{w}}\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = \mathbb{E}_{y,\boldsymbol{X}}\big[(y - f_{\boldsymbol{\theta}}(\boldsymbol{X}))\nabla_{\boldsymbol{w}}f_{\boldsymbol{\theta}}(\boldsymbol{X})\big]$.

The first gradient step is easy to calculate and returns a classifier estimate that is already in the direction of $\boldsymbol{w}_\star$.

Lemma 1 (Population first step)**.** *The first population gradient step* $\boldsymbol{w}_1 = \eta \boldsymbol{G}_{\boldsymbol{w}}(0,0)$ *satisfies* $\boldsymbol{w}_1 = \eta\zeta \boldsymbol{w}_\star$ *since under* (DATA)*,* $\boldsymbol{G}_{\boldsymbol{w}}(0,0) = \frac{\mathbb{E}_{y,\boldsymbol{x}}[y\boldsymbol{X}^\top \mathbb{1}]}{T} = \zeta \boldsymbol{w}_\star$.

The second gradient step $\boldsymbol{q}_1 = \gamma \boldsymbol{G}_{\boldsymbol{w}}(\boldsymbol{w}_1, 0)$ is more intricate: unless $\boldsymbol{q}_\star \perp \boldsymbol{w}_\star$, $\boldsymbol{q}_1$ also has nonzero components in both directions $\boldsymbol{q}_\star$ and $\boldsymbol{w}_\star$.

Lemma 2 (Population second step)**.** *The second population gradient step* $\boldsymbol{q}_1 = \gamma \boldsymbol{G}_{\boldsymbol{w}}(\boldsymbol{w}_1, 0)$ *satisfies for* $\alpha := \eta\zeta$

$$\begin{aligned} \boldsymbol{q}_1 = \gamma\alpha W^2(\zeta - \zeta^2)\big(1 + \frac{\alpha\sigma^2}{T} - \alpha\zeta(W^2 + \rho^2 Q^2)\big)\boldsymbol{q}_\star \\ + \gamma\alpha\rho Q W(\zeta - \zeta^2)\big(1 - 2\zeta\alpha W^2 - (1 + \frac{2}{T})\alpha\sigma^2\big)\,\boldsymbol{w}_\star\,. \end{aligned} \tag{4}$$

Proof. Since this computation involves several terms, we defer complete proof to Appendix G.1. The above simplification is made possible by leveraging the third-moment noise Assumption 1.a. □

Lemma 2 highlights the following key aspects: (i) As mentioned, $\boldsymbol{q}_1$ also picks up the $\boldsymbol{w}_\star$ direction unless $\rho = 0$. However, we can control the magnitude of this undesired term by choosing small step-size η (see Cor. 1). (ii) As αW^2 grows, the gradient component in the $\boldsymbol{q}_\star$ direction might end up pointing in the direction of $-\boldsymbol{q}_\star$. This is because large signal along the $\boldsymbol{w}_\star$ direction might still allow to predict ± 1 label. However, this can always be avoided by choosing sufficiently small step-size η (see Cor. 1). (iii) Relatedely, as the noise strength σ^2 grows, gradient in the $\boldsymbol{q}_\star$ direction grows as well. This is because, going along $\boldsymbol{q}_\star$ direction attenuates the noise and cleans up the prediction. (iv) Finally, as $\zeta \to 1$ and $\zeta - \zeta^2 \to 0$ the magnitude of the gradient decays because all tokens contain signal information and there is no need for $\boldsymbol{q}_\star$.

To see how $\boldsymbol{q}_1$ selects good tokens, we investigate the relevance scores (normalized by the step size γ) $r_t := \boldsymbol{x}_t^\top \boldsymbol{q}_1/\gamma$ of signal vs noise tokens. Attending to context-relevant tokens requires the signal/context relevance scores be larger than the noisy ones. Concretely, suppose we have

$$B := \min_{t\in\mathcal{R}}\{r_t = \frac{(\boldsymbol{q}_\star + y\boldsymbol{w}_\star^\top)\boldsymbol{q}_1}{\gamma}\} \geq 2 \max_{t\in\mathcal{R}^c}\{r_t = \frac{\boldsymbol{z}_t^\top \boldsymbol{q}_1}{\gamma}\}. \tag{5}$$

Then, $|\mathcal{R}|\, e^{\gamma B} + |\mathcal{R}^c|\, e^{\gamma B/2} \geq S := \sum_{t'\in[T]} e^{\gamma r_{t'}} \geq |\mathcal{R}|\, e^{\gamma B}$, which implies the following for the attention weights as step size increases $\gamma \to \infty$:

$$a_t = [\phi(\boldsymbol{X}\boldsymbol{q}_1)]_t = e^{\gamma r_t}/S \begin{cases} = \frac{e^{\gamma B}}{S} \to \frac{1}{\zeta T} & t \in \mathcal{R} \\ \leq \frac{e^{\gamma B/2}}{S} \to 0 & t \in \mathcal{R}^c \end{cases}. \tag{6}$$

Provided (5) holds, a large enough second gradient step (large γ) finds $\boldsymbol{q}_1$ that attends (nearly) perfectly to context-relevant tokens in $\mathcal{R}$ and attenuates (almost) all the noise tokens in $\mathcal{R}^c$. The following theorem formalizes the above intuition. We defer the complete proof to Appendix G.3.

Theorem 5 (Tying things together)**.** *Consider the model* $\boldsymbol{\theta}^\gamma = (\boldsymbol{w}_2^\gamma, \boldsymbol{q}_1^\gamma)$ *where* $\boldsymbol{q}_1^\gamma = \gamma \boldsymbol{G}_{\boldsymbol{q}}(\boldsymbol{w}_1, 0)$*,* $\boldsymbol{w}_2^\gamma = \boldsymbol{G}_{\boldsymbol{w}}(0, \boldsymbol{q}_1^\gamma)$ *and* $\boldsymbol{w}_1 = \eta \boldsymbol{G}_{\boldsymbol{w}}(0,0)$ *for step-size* η *small enough (see Eq. (58) for details). Then, there exists absolute constant* $C > 0$ *and sufficiently large* $\alpha > 1$ *and step-size* γ *such that* $\mathrm{ERR}\left(f_{\boldsymbol{\theta}^\gamma}\right) \leq 2/T^{\alpha-1}$*, provided*

$$\sigma\sqrt{\log\left(2(1-\zeta)T^\alpha\right)} \leq \frac{(1-\rho^2/2)Q - 2|\rho|W}{2\sqrt{2}\sqrt{1+3\rho^2}}. \tag{7}$$

Eq. (7) guarantees the desired condition (5) holds. When $\rho = 0$ ($\boldsymbol{q}_\star \perp \boldsymbol{w}_\star$), it requires $Q = \Omega(\sqrt{\log(T)})$ to attend to context-relevant tokens irrespective of sparsity level ζ. For $\rho \neq 0$, (7) further imposes $|\rho| < \frac{Q}{2W}$. Here the role of Q, W is reversed compared to $|\rho| \leq \frac{W}{Q}$ in 3.a. The latter guarantees classifier energy is larger so that signal $y\boldsymbol{w}_\star$ dominates $\boldsymbol{q}_\star$, while for prompt-attention to attend to relevant tokens it is favorable that energy of $\boldsymbol{q}_\star$ dominates $\boldsymbol{w}_\star$.

Finally, we compare the theorem's error to the error $Q(\sqrt{\frac{\zeta^2 W^2 T}{1-\zeta}})$ of the linear model in Fact D.1. For concreteness, fix $W = O(1)$, $Q = O(\log(T))$ (satisfying (7)) and extreme sparsity $\zeta = O(1/\sqrt{T})$. Then the error of linear model is $O(1)$, while the error $O(1/T^{\alpha-1})$ of (population) Algorithm 3 for prompt-attention is decreasing in T.

E.3 Finite-sample analysis

Here, we investigate the behavior of Algorithm 3 with finite sample-size n. For convenience, we first introduce an additional de-biasing step after calculating the three gradients in (3). Specifically, for a sample $\mathcal{S}_4$ of size n we compute a bias variable $\widehat{b} := \frac{1}{n}\sum_{i=1}^{n} f_{(\widehat{\boldsymbol{q}}_1, \widehat{\boldsymbol{w}}_2)}(\boldsymbol{X}_i)$, and use it to de-bias the model's prediction by outputting $f_{(\boldsymbol{\theta},b)}(\boldsymbol{X}) := f_{\boldsymbol{\theta}}(\boldsymbol{X}) - b$. While this extra step is not necessary it simplifies the statement of our results. Intuitively, $\widehat{b}$ helps with adjusting the decision boundary by removing contributions of the context vector in the final prediction (the context vector is useful only for token-selection rather than final prediction).

Below we provide a simplified version of our main result. Refer to Theorem 7 for precise details. For clarity, we fix noise variance $\sigma \propto 1$ and use $\gtrsim, \tilde{\mathcal{O}}()$ to suppress logarithmic dependencies on T, n.

Theorem 6 (Error rate)**.** *Consider $\boldsymbol{\theta} = (\widehat{\boldsymbol{w}}_2, \widehat{\boldsymbol{q}}_1)$ as per Algorithm 3 with bias $\widehat{b}$ as explained above. Choose step sizes $\eta \propto \frac{1}{Q^2\zeta}$ and γ large enough. Suppose $Q \gtrsim 1 + W$, $\zeta \le 0.9$. Declare $\texttt{rate}_{\text{LIN}} = \zeta^2 W^2 T$ and set error rate to*

$$\texttt{rate} := Q^2 \wedge Q\sqrt{d} \wedge n^{1/3}\zeta^{2/3}(W/Q)^{4/3} \wedge (n/d)\texttt{rate}_{\text{LIN}}.$$

For a constant $c > 0$, with probability $1 - ce^{-\tilde{\mathcal{O}}(\texttt{rate}\wedge d)}$ over the training process, the classification error obeys

$$\text{ERR}(f_{(\boldsymbol{\theta},\widehat{b})}) \le ce^{-\tilde{\mathcal{O}}(\texttt{rate})}.$$

A few remarks are in place. $\texttt{rate}_{\text{LIN}}$ corresponds to the linear baseline f^{LIN} as it is known that (see Fact D.1) its error rate (under gaussian noise) follows $\texttt{rate}_{\text{LIN}}$. This rate appears in our theorem with $(n/d)\texttt{rate}_{\text{LIN}}$; thus, as soon as $n \propto d$, gradient-descent can (potentially) beat the linear model. The remaining components of our $\texttt{rate}$ are regularity conditions: Importantly, we ask for Q to be large enough so that the attention weights $\phi(\boldsymbol{X}\hat{\boldsymbol{q}}_1)$ provably select the relevant tokens and discard those irrelevant. This is analogous to its population-counterpart condition (7). We also need $n \gtrsim (W/Q)^4\zeta^2$ to suppress the variance due to the $\boldsymbol{q}_\star$ terms during perturbation analysis of $\hat{\boldsymbol{w}}_1$. Observe that, by letting $n \to \infty$ in Theorem 6, we end up with $\texttt{rate} = Q(Q \wedge \sqrt{d})$ to obtain an error rate of $e^{-cQ(Q\wedge\sqrt{d})}$. This is a very distinct learning regime compared to f^{LIN} and provably beats it whenever $Q(Q \wedge \sqrt{d}) \gtrsim \zeta^2 W^2 T$. This conclusion is consistent with the discussion following Theorem 5.

Sharp error rates: Finally, in Appendix D we provide an exact analysis of the classification error when $\boldsymbol{q}_\star$ is known and only $\boldsymbol{w}_\star$ is estimated from data. This analysis exactly quantifies the value of context-information and how prompt-tuning retrieves it. Specifically, we prove a sharp asymptotic error rate of $\mathcal{Q}\left(\frac{e^{Q^2/4}}{\sqrt{1+\texttt{ISNR}(n/d)}} \cdot \sqrt{\frac{\zeta^2 W^2 T}{1-\zeta}}\right)$ where $\texttt{ISNR}(\alpha) := \alpha^{-1}\frac{(1-\zeta)e^{-Q^2/2}}{\texttt{rate}_{\text{LIN}}}$, $\mathcal{Q}(\cdot)$ is the gaussian tail function and noise is isotropic gaussian. This strictly improves over the existing optimal rates for (context-free) Gaussian mixture models thanks to the context information.

F Proofs for finite-sample Gradient Analysis in Section E.3

In this section, we focus on finite-sample analysis of Algorithm 3. Introduce the following shorthand notation analogous to the population counterparts in Section E.2:

$$\widehat{\boldsymbol{G}}_{\boldsymbol{q}}(\boldsymbol{q},\boldsymbol{w}) := -\nabla_{\boldsymbol{q}}\widehat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i\in[n]}(y_i - f_{\boldsymbol{\theta}}(\boldsymbol{X}_i))\boldsymbol{X}_i^\top \phi'(\boldsymbol{X}_i\boldsymbol{q})\,\boldsymbol{X}_i\boldsymbol{w}$$

$$\widehat{\boldsymbol{G}}_{\boldsymbol{w}}(\boldsymbol{q},\boldsymbol{w}) := -\nabla_{\boldsymbol{w}}\widehat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i\in[n]}(y_i - f_{\boldsymbol{\theta}}(\boldsymbol{X}_i))\boldsymbol{X}_i^\top \phi((\boldsymbol{X}_i\boldsymbol{q})). \tag{8}$$

F.1 Gradient Calculations

We begin with the gradient calculations for the first two steps of the algorithm.

For convenience, we make use of the following shorthands

$$R_{\boldsymbol{q}_\star} := R_{\boldsymbol{w},\boldsymbol{q}_\star} := \boldsymbol{w}^\top \boldsymbol{q}_\star, \qquad R_{\boldsymbol{w}_\star} := R_{\boldsymbol{w},\boldsymbol{w}_\star} := \boldsymbol{w}^\top \boldsymbol{w}_\star, \qquad \alpha_i := \alpha(\boldsymbol{w}, y_i) := R_{\boldsymbol{q}_\star} + y_i R_{\boldsymbol{w}_\star},$$
$$\boldsymbol{\gamma}_i := \boldsymbol{\gamma}(\boldsymbol{Z}_i) = \frac{1}{T}\boldsymbol{Z}_i^T \mathbb{1}, \qquad \beta_i := \beta(\boldsymbol{Z}_i; \boldsymbol{w}) = \boldsymbol{\gamma}_i^T \boldsymbol{w}, \qquad \hat{\boldsymbol{\Sigma}}_i := \frac{1}{T}\boldsymbol{Z}_i^T \boldsymbol{Z}_i.$$

for $\boldsymbol{Z}_i \in \mathbb{R}^{(1-\zeta)T \times d}$ is the matrix of irrelevant tokens $\boldsymbol{z}_{i,t}, t \in \mathcal{R}^c$ for sample $i \in [n]$.

Lemma 3. *Under dataset model (DATA) and Assumption 1.a, we have*

$$\widehat{\boldsymbol{G}}_{\boldsymbol{w}}(0,0) = \zeta \boldsymbol{w}_\star + \zeta \boldsymbol{q}_\star \left(\frac{1}{n} \sum_{i \in [n]} y_i \right) + \frac{1}{n} \sum_i y_i \boldsymbol{\gamma}_i. \tag{10}$$

Lemma 4. *Under dataset model (DATA) and Assumption 1.a, we have that*

$$\widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0,\boldsymbol{w}) := \frac{1}{n} \sum_{i=1}^n (y_i - \zeta\alpha_i - \beta_i) \left[\left((\zeta - \zeta^2)\alpha_i - \zeta\beta_i \right) (\boldsymbol{q}_\star + y_i \boldsymbol{w}_\star) + \hat{\boldsymbol{\Sigma}}_i \boldsymbol{w} - (\zeta\alpha_i + \beta_i)\boldsymbol{\gamma}_i \right]. \tag{11}$$

F.1.1 Proof of Lemma 3

By direct computation,

$$\begin{aligned}
\hat{\boldsymbol{G}}_{\boldsymbol{w}}(0,0) := -\nabla_{\boldsymbol{w}} \widehat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{\theta}) &= \frac{1}{nT} \sum_{i \in [n]} y_i \boldsymbol{X}_i^T \mathbb{1}_T = \frac{1}{nT} \sum_{i \in [n]} \sum_{t \in [T]} y_i \boldsymbol{x}_{i,t} \\
&= \frac{1}{n} \Big(\sum_{i \in [n]} \zeta_i y_i \Big) \boldsymbol{q}_\star + \frac{1}{n} \Big(\sum_{i \in [n]} \zeta_i \Big) \boldsymbol{w}_\star + \frac{1}{n} \sum_{i \in [n]} y_i \boldsymbol{z}_{\text{avg},i}
\end{aligned}$$

where ζ_i is the fraction of relevant tokens in the i-th sample and $\boldsymbol{z}_{\text{avg},i} = \frac{1}{T}\sum_{t=1}^T \boldsymbol{z}_{i,t}$ where we again set noise $\boldsymbol{z}_t = 0$ for relevant tokens.

F.1.2 Proof of Lemma 4

Recall that $\phi'(0) = \frac{1}{T}\mathbb{I} - \frac{1}{T^2}\mathbb{1}\mathbb{1}^T$; hence, for $\boldsymbol{\theta} = (0, \boldsymbol{w})$:

$$\hat{\boldsymbol{G}}_{\boldsymbol{q}}(0,\boldsymbol{w}) := -\nabla_{\boldsymbol{q}} \widehat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i \in [n]} \underbrace{\frac{1}{T}(y_i - f_{\boldsymbol{\theta}}(\boldsymbol{X}_i)) \boldsymbol{X}_i^\top \boldsymbol{X}_i \boldsymbol{w}}_{\text{Term}_{1,i}} - \frac{1}{n} \sum_{i \in [n]} \underbrace{\frac{1}{T^2}(y_i - f_{\boldsymbol{\theta}}(\boldsymbol{X}_i)) \boldsymbol{X}_i^\top \mathbb{1}\mathbb{1}^\top \boldsymbol{X}_i \boldsymbol{w}}_{\text{Term}_{2,i}}.$$

Moreover, note that,

$$\begin{aligned}
f_{\boldsymbol{\theta}}(\boldsymbol{X}_i) &= \frac{1}{T} \boldsymbol{w}^T \boldsymbol{X}_i^T \mathbb{1}, \\
\boldsymbol{X}_i^\top \boldsymbol{X}_i &= \zeta T (\boldsymbol{q}_\star + y_i \boldsymbol{w}_\star)(\boldsymbol{q}_\star + y_i \boldsymbol{w}_\star)^\top + \boldsymbol{Z}_i^\top \boldsymbol{Z}_i, \\
\boldsymbol{X}_i^\top \mathbb{1} &= \zeta T (\boldsymbol{q}_\star + y_i \boldsymbol{w}_\star) + \boldsymbol{Z}_i^\top \mathbb{1}.
\end{aligned}$$

where recall the notation in Lemma 3 for $\boldsymbol{Z}_i$. Hence, using the lemma's notation (repeated here for convenience)

$$\alpha_i := R_{\boldsymbol{q}_\star} + y_i R_{\boldsymbol{w}_\star}, \quad \beta_i := \beta(\boldsymbol{Z}_i; \boldsymbol{w}) := \frac{1}{T} \mathbb{1}^T \boldsymbol{Z}_i \boldsymbol{w}, \quad \boldsymbol{\gamma}_i := \boldsymbol{\gamma}(\boldsymbol{Z}_i) = \frac{1}{T} \boldsymbol{Z}_i^T \mathbb{1}, \quad \hat{\boldsymbol{\Sigma}}_i := \frac{1}{T} \boldsymbol{Z}_i^T \boldsymbol{Z}_i.$$

we find that

$$y_i - f_{\boldsymbol{\theta}}(\boldsymbol{X}_i) = y_i - \zeta\alpha_i - \beta_i \tag{12}$$

$$\frac{1}{T} \boldsymbol{X}_i \boldsymbol{X}_i^\top \boldsymbol{w} = \zeta\alpha_i \boldsymbol{q}_\star + \zeta y_i \alpha_i \boldsymbol{w}_\star + \hat{\boldsymbol{\Sigma}}_i \boldsymbol{w} \tag{13}$$

$$\frac{1}{T^2} \boldsymbol{X}_i^\top \mathbb{1}\mathbb{1}^\top \boldsymbol{X}_i \boldsymbol{w} = \zeta \left(\zeta\alpha_i + \beta_i \right) \boldsymbol{q}_\star + \zeta \left(\zeta\alpha_i + \beta_i \right) y_i \boldsymbol{w}_\star + \left(\zeta\alpha_i + \beta_i \right) \boldsymbol{\gamma}_i. \tag{14}$$

With the above, each one of the two terms becomes:

$$\text{Term}_{1,\text{i}} = (y_i - \zeta\alpha_i - \beta_i)\,\zeta\alpha_i \boldsymbol{q}_\star + (y_i - \zeta\alpha_i - \beta_i)\,\zeta\alpha_i y_i \boldsymbol{w}_\star + (y_i - \zeta\alpha_i - \beta_i)\,\hat{\boldsymbol{\Sigma}}_i \boldsymbol{w} \tag{15}$$

$$\text{Term}_{2,\text{i}} = \zeta\,(y_i - \zeta\alpha_i - \beta_i)\,(\zeta\alpha_i + \beta_i)\,\boldsymbol{q}_\star + \zeta\,(y_i - \zeta\alpha_i - \beta_i)\,(\zeta\alpha_i + \beta_i)\,y_i \boldsymbol{w}_\star + (y_i - \zeta\alpha_i - \beta_i)\,(\zeta\alpha_i + \beta_i)\,\boldsymbol{\gamma}_i \tag{16}$$

Combining the above:

$$\text{Term}_{1,\text{i}} - \text{Term}_{2,\text{i}} = (y_i - \zeta\alpha_i - \beta_i)\left[\zeta\left((1-\zeta)\alpha_i - \beta_i\right)(\boldsymbol{q}_\star + y_i \boldsymbol{w}_\star) + \hat{\boldsymbol{\Sigma}}_i \boldsymbol{w} - (\zeta\alpha_i + \beta_i)\boldsymbol{\gamma}_i\right] \tag{17}$$

F.2 Concentration of Gradient $\widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0, \boldsymbol{w})$ in the $\boldsymbol{q}$ direction

The main result of this section is the following lemma about concentration of gradient with respect to $\boldsymbol{q}$.

Lemma 5 (Concentration of $\widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0, \boldsymbol{w})$)**.** *Fix any vectors $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^d$. For convenience define $R_{\boldsymbol{v},\boldsymbol{q}_\star} := \boldsymbol{v}^T \boldsymbol{q}_\star$ and $R_{\boldsymbol{v},\boldsymbol{w}_\star} := \boldsymbol{v}^T \boldsymbol{w}_\star$ and recall $R_{\boldsymbol{w}_\star}, R_{\boldsymbol{q}_\star}$ notations from Lemma 4. Then, we can decompose*

$$\boldsymbol{v}^T \widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0, \boldsymbol{w}) = \boldsymbol{v}^T \boldsymbol{G}_{\boldsymbol{q}}(0, \boldsymbol{w}) + \boldsymbol{v}^T \widetilde{\boldsymbol{G}}_{\boldsymbol{q}}(0, \boldsymbol{w}),$$

where the expectation term is given by

$$\begin{aligned}\boldsymbol{v}^T \boldsymbol{G}_{\boldsymbol{q}}(0, \boldsymbol{w}) := \mathbb{E}\left[\boldsymbol{v}^T \widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0, \boldsymbol{w})\right] = &\left((\zeta - \zeta^2)\left(R_{\boldsymbol{w}_\star} + \boldsymbol{w}^T \boldsymbol{\Sigma} \boldsymbol{w}/T\right) - (\zeta^2 - \zeta^3)\left(R^2_{\boldsymbol{w}_\star} + R^2_{\boldsymbol{q}_\star}\right)\right) R_{\boldsymbol{v},\boldsymbol{q}_\star} \\ &+ \left(\left((\zeta - \zeta^2) - 2(\zeta^2 - \zeta^3) R_{\boldsymbol{w}_\star}\right) R_{\boldsymbol{q}_\star}\right) R_{\boldsymbol{w},\boldsymbol{q}_\star} - \left((1 + 2/T)\,(\zeta - \zeta^2) R_{\boldsymbol{q}_\star}\right) \boldsymbol{v}^T \boldsymbol{\Sigma} \boldsymbol{w}\end{aligned} \tag{18}$$

and the deviation term obeys

$$\begin{aligned}
\boldsymbol{v}^T\widetilde{\boldsymbol{G}}_{\boldsymbol{q}}(0,\boldsymbol{w}) = &\left[\left((\zeta-\zeta^2)-2(\zeta^2-\zeta^3)R_{\boldsymbol{w}_\star}\right)R_{\boldsymbol{q}_\star}R_{\boldsymbol{v},\boldsymbol{q}_\star}+\left((\zeta-\zeta^2)R_{\boldsymbol{w}_\star}-(\zeta^2-\zeta^3)\left(R^2_{\boldsymbol{w}_\star}+R^2_{\boldsymbol{q}_\star}\right)\right)R_{\boldsymbol{v},\boldsymbol{w}_\star}\right]\left(\frac{1}{n}\sum_{i=1}^n y_i\right)\\
&+\left[\left(-\zeta+2\zeta^2\right)R_{\boldsymbol{q}_\star}R_{\boldsymbol{v},\boldsymbol{q}_\star}+\left(-\zeta+(-\zeta+2\zeta^2)R_{\boldsymbol{w}_\star}\right)R_{\boldsymbol{v},\boldsymbol{w}_\star}+(1-\zeta)\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{v}\right]\left(\frac{1}{n}\sum_{i=1}^n\left(\boldsymbol{\gamma}_i^T\boldsymbol{w}\right)\right)\\
&+\left[\zeta R_{\boldsymbol{w}_\star}-\zeta^2(R^2_{\boldsymbol{q}_\star}+R^2_{\boldsymbol{w}_\star})+\frac{(1-\zeta)}{T}\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}\right]\left(\frac{1}{n}\sum_{i=1}^n\boldsymbol{v}^T\boldsymbol{\gamma}_i\right)\\
&+\left[\left(-\zeta+(-\zeta+2\zeta^2)R_{\boldsymbol{w}_\star}\right)R_{\boldsymbol{v},\boldsymbol{q}_\star}+\left(-\zeta+2\zeta^2\right)R_{\boldsymbol{q}_\star}R_{\boldsymbol{v},\boldsymbol{w}_\star}\right]\left(\frac{1}{n}\sum_{i=1}^n y_i\left(\boldsymbol{\gamma}_i^T\boldsymbol{w}\right)\right)\\
&+\left[\zeta R_{\boldsymbol{q}_\star}-2\zeta^2R_{\boldsymbol{q}_\star}R_{\boldsymbol{w}_\star}\right]\left(\frac{1}{n}\sum_{i=1}^n y_i\left(\boldsymbol{v}^T\boldsymbol{\gamma}_i\right)\right)\\
&+\zeta R_{\boldsymbol{v},\boldsymbol{q}_\star}\left(\frac{1}{n}\sum_{i=1}^n\left(\boldsymbol{\gamma}_i^T\boldsymbol{w}\right)^2-\frac{(1-\zeta)}{T}\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}\right)\\
&+\zeta R_{\boldsymbol{v},\boldsymbol{w}_\star}\left(\frac{1}{n}\sum_{i=1}^n\left(\boldsymbol{\gamma}_i^T\boldsymbol{w}\right)^2y_i\right)\\
&+\left[1-2\zeta R_{\boldsymbol{w}_\star}\right]\left(\frac{1}{n}\sum_{i=1}^n y_i\left(\boldsymbol{w}^T\boldsymbol{\gamma}_i\right)\left(\boldsymbol{v}^T\boldsymbol{\gamma}_i\right)\right)\\
&+(1-\zeta R_{\boldsymbol{w}_\star})\left(\frac{1}{n}\sum_{i=1}^n y_i\boldsymbol{v}^T\hat{\boldsymbol{\Sigma}}_i\boldsymbol{w}\right)\\
&-\zeta R_{\boldsymbol{q}_\star}\left(\frac{1}{n}\sum_{i=1}^n\boldsymbol{v}^T\hat{\boldsymbol{\Sigma}}_i\boldsymbol{w}-(1-\zeta)\boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{w}\right)\\
&-\left[2\zeta R_{\boldsymbol{q}_\star}\right]\left(\frac{1}{n}\sum_{i=1}^n\left(\boldsymbol{w}^T\boldsymbol{\gamma}_i\right)\left(\boldsymbol{v}^T\boldsymbol{\gamma}_i\right)-\frac{1-\zeta}{T}\boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{w}\right)\\
&-\left(\frac{1}{n}\sum_{i=1}^n\left(\boldsymbol{v}^T\boldsymbol{\gamma}_i\right)\left(\left(\boldsymbol{w}^T\boldsymbol{\gamma}_i\right)^2-\frac{(1-\zeta)}{T}\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}\right)\right)\\
&-\left(\frac{1}{n}\sum_{i=1}^n\left(\boldsymbol{\gamma}_i^T\boldsymbol{w}\right)\left(\boldsymbol{v}^T\hat{\boldsymbol{\Sigma}}_i\boldsymbol{w}-(1-\zeta)\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{v}\right)\right). \qquad (19)
\end{aligned}$$

Moreover, all random terms in (19) are zero-mean and concentrate as prescribed by Lemma 6 below

Lemma 6 (Main concentration lemma)**.** *Let $y_i, i\in[n]$ be iid Rademacher random variables. Let $\boldsymbol{Z}_i\in\mathbb{R}^{(1-\zeta)T\times d}, i\in[n]$ be iid copies of a random matrix $\boldsymbol{Z}$. Each row $\boldsymbol{z}_t, t\in[(1-\zeta)T]$ of $\boldsymbol{Z}$ is an iid copy of a random vector $\boldsymbol{z}$ satisfying Assumption 1.a. For convenience denote $\boldsymbol{\gamma}_i := \boldsymbol{Z}_i^T\mathbb{1}/T$ and $\hat{\boldsymbol{\Sigma}}_i := \boldsymbol{Z}_i^T\boldsymbol{Z}_i/T$. Then, the following statements are true for all vectors $\boldsymbol{w},\boldsymbol{v}\in\mathbb{R}^d$.*

$$\left\|\frac{1}{n}\sum_{i\in[n]}y_i\right\|_{\psi_2}\le\frac{C}{\sqrt{n}}$$

$$\left\|\frac{1}{n}\sum_{i\in[n]}\boldsymbol{\gamma}_i^T\boldsymbol{w}\right\|_{\psi_2}\vee\left\|\frac{1}{n}\sum_{i\in[n]}y_i\boldsymbol{\gamma}_i^T\boldsymbol{w}\right\|_{\psi_2}\le\frac{C\sigma\sqrt{1-\zeta}\|\boldsymbol{w}\|_2}{\sqrt{nT}}$$

$$\left\|\frac{1}{n}\sum_{i\in[n]}\left(\boldsymbol{\gamma}_i^T\boldsymbol{w}\right)\left(\boldsymbol{\gamma}_i^T\boldsymbol{v}\right)-\frac{1-\zeta}{T}\boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{w}\right\|_{\psi_1}\vee\left\|\frac{1}{n}\sum_{i\in[n]}y_i\left(\boldsymbol{\gamma}_i^T\boldsymbol{w}\right)\left(\boldsymbol{\gamma}_i^T\boldsymbol{v}\right)\right\|_{\psi_1}\le\frac{C\sigma^2(1-\zeta)\|\boldsymbol{w}\|_2\|\boldsymbol{v}\|_2}{T\sqrt{n}}$$

$$\left\|\frac{1}{n}\sum_{i\in[n]}\boldsymbol{w}^T\hat{\boldsymbol{\Sigma}}_i\boldsymbol{v}-(1-\zeta)\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{v}\right\|_{\psi_1}\vee\left\|\frac{1}{n}\sum_{i\in[n]}y_i\boldsymbol{w}^T\hat{\boldsymbol{\Sigma}}_i\boldsymbol{v}\right\|_{\psi_1}\le\frac{C\sigma^2\sqrt{1-\zeta}\|\boldsymbol{w}\|_2\|\boldsymbol{v}\|_2}{\sqrt{nT}}$$

$$\left\| \frac{1}{n} \sum_{i \in [n]} \left(\left(\boldsymbol{\gamma}_i^T \boldsymbol{w} \right)^2 - \frac{1-\zeta}{T} \boldsymbol{w}^T \boldsymbol{\Sigma} \boldsymbol{w} \right) \boldsymbol{\gamma}_i^T \boldsymbol{v} \right\|_{\psi_{2/3}} \leq \frac{C \sigma^3 (1-\zeta)^{3/2} \|\boldsymbol{w}\|_2^2 \|\boldsymbol{v}\|_2}{T^{3/2} \sqrt{n}} \log n$$

$$\left\| \frac{1}{n} \sum_{i \in [n]} \left(\boldsymbol{\gamma}_i^T \boldsymbol{w} \right) \left(\boldsymbol{w}^T \hat{\boldsymbol{\Sigma}}_i \boldsymbol{v} - (1-\zeta) \boldsymbol{w}^T \boldsymbol{\Sigma} \boldsymbol{v} \right) \right\|_{\psi_{2/3}} \leq \frac{C K^3 (1-\zeta) \|\boldsymbol{w}\|_2^2 \|\boldsymbol{v}\|_2}{T \sqrt{n}} \log n$$

Also, all the random variables that appear above are zero mean.

F.2.1 Proof of Lemma 5

We split (11) in four terms and handle each of them separately.

- **Term**$_{\mathrm{I}} = \frac{1}{n} \sum_{i=1}^n \left(y_i - \zeta \alpha_i - \beta_i \right) \left((\zeta - \zeta^2) \alpha_i - \zeta \beta_i \right) \boldsymbol{q}_\star$

We first focus on

$$\mathrm{Term}_{\mathrm{I}} = \frac{1}{n} \sum_{i=1}^n \left(y_i (1 - \zeta R_{\boldsymbol{w}_\star}) - \beta_i - \zeta R_{\boldsymbol{q}_\star} \right) \left(y_i (\zeta - \zeta^2) R_{\boldsymbol{w}_\star} - \zeta \beta_i + (\zeta - \zeta^2) R_{\boldsymbol{q}_\star} \right) \boldsymbol{q}_\star =: A \boldsymbol{q}_\star .$$

We we can express A above conveniently as follows (recall $y_i^2 = 1$):

$$\begin{aligned}
A \coloneqq & -\zeta(\zeta - \zeta^2) R_{\boldsymbol{q}_\star}^2 + (1 - \zeta R_{\boldsymbol{w}_\star})(\zeta - \zeta^2) R_{\boldsymbol{w}_\star} + \left((1 - \zeta R_{\boldsymbol{w}_\star})(\zeta - \zeta^2) R_{\boldsymbol{q}_\star} - \zeta(\zeta - \zeta^2) R_{\boldsymbol{w}_\star} R_{\boldsymbol{q}_\star} \right) \left(\frac{1}{n} \sum_{i=1}^n y_i \right) \\
& + \left(-(\zeta - \zeta^2) R_{\boldsymbol{q}_\star} + \zeta^2 R_{\boldsymbol{q}_\star} \right) \left(\frac{1}{n} \sum_{i=1}^n \beta_i \right) + \left(-(1 - \zeta R_{\boldsymbol{w}_\star}) \zeta - (\zeta - \zeta^2) R_{\boldsymbol{w}_\star} \right) \left(\frac{1}{n} \sum_{i=1}^n y_i \beta_i \right) + \zeta \left(\frac{1}{n} \sum_{i=1}^n \beta_i^2 \right) \\
= & (\zeta - \zeta^2) R_{\boldsymbol{w}_\star} - (\zeta^2 - \zeta^3) \left(R_{\boldsymbol{w}_\star}^2 + R_{\boldsymbol{q}_\star}^2 \right) + \left((\zeta - \zeta^2) - 2(\zeta^2 - \zeta^3) R_{\boldsymbol{w}_\star} \right) R_{\boldsymbol{q}_\star} \left(\frac{1}{n} \sum_{i=1}^n y_i \right) \\
& + \left(-\zeta + 2\zeta^2 \right) R_{\boldsymbol{q}_\star} \left(\frac{1}{n} \sum_{i=1}^n \beta_i \right) + \left(-\zeta + (-\zeta + 2\zeta^2) R_{\boldsymbol{w}_\star} \right) \left(\frac{1}{n} \sum_{i=1}^n y_i \beta_i \right) \\
& + \zeta \left(\frac{1}{n} \sum_{i=1}^n \beta_i^2 - \frac{(1-\zeta)}{T} \boldsymbol{w}^T \boldsymbol{\Sigma} \boldsymbol{w} \right) + \frac{(\zeta - \zeta^2)}{T} \boldsymbol{w}^T \boldsymbol{\Sigma} \boldsymbol{w}.
\end{aligned}$$

From Lemma 6, all random terms above are zero mean. Hence,

$$\begin{aligned}
\mathbb{E}[A] &= -(\zeta^2 - \zeta^3) R_{\boldsymbol{q}_\star}^2 + (1 - \zeta R_{\boldsymbol{w}_\star})(\zeta - \zeta^2) R_{\boldsymbol{w}_\star} + (\zeta - \zeta^2) \frac{\boldsymbol{w}^T \boldsymbol{\Sigma} \boldsymbol{w}}{T} \\
&= -(\zeta^2 - \zeta^3) R_{\boldsymbol{q}_\star}^2 + (\zeta - \zeta^2) \left(R_{\boldsymbol{w}_\star} + \boldsymbol{w}^T \boldsymbol{\Sigma} \boldsymbol{w} / T \right) - (\zeta^2 - \zeta^3) R_{\boldsymbol{w}_\star}^2 \\
&= (\zeta - \zeta^2) \left(R_{\boldsymbol{w}_\star} + \boldsymbol{w}^T \boldsymbol{\Sigma} \boldsymbol{w} / T \right) - (\zeta^2 - \zeta^3) \left(R_{\boldsymbol{w}_\star}^2 + R_{\boldsymbol{q}_\star}^2 \right).
\end{aligned} \tag{20}$$

- **Term**$_{\mathrm{II}} = \frac{1}{n} \sum_{i=1}^n \left(y_i - \zeta \alpha_i - \beta_i \right) \left((\zeta - \zeta^2) \alpha_i - \zeta \beta_i \right) y_i \boldsymbol{w}_\star$

$$\mathrm{Term}_{\mathrm{II}} = \frac{1}{n} \sum_{i=1}^n \left(y_i (1 - \zeta R_{\boldsymbol{w}_\star}) - \beta_i - \zeta R_{\boldsymbol{q}_\star} \right) \left(y_i (\zeta - \zeta^2) R_{\boldsymbol{w}_\star} - \zeta \beta_i + (\zeta - \zeta^2) R_{\boldsymbol{q}_\star} \right) y_i \boldsymbol{w}_\star = B \boldsymbol{w}_\star .$$

We we can express B above conveniently as:

$$\begin{aligned}
B \coloneqq & \left((\zeta - \zeta^2) R_{\boldsymbol{w}_\star} - (\zeta^2 - \zeta^3) \left(R_{\boldsymbol{w}_\star}^2 + R_{\boldsymbol{q}_\star}^2 \right) \right) \left(\frac{1}{n} \sum_{i \in [n]} y_i \right) + \left((\zeta - \zeta^2) - 2(\zeta^2 - \zeta^3) R_{\boldsymbol{w}_\star} \right) R_{\boldsymbol{q}_\star} \\
& + \left(-\zeta + 2\zeta^2 \right) R_{\boldsymbol{q}_\star} \left(\frac{1}{n} \sum_{i=1}^n \beta_i y_i \right) + \left(-\zeta + (-\zeta + 2\zeta^2) R_{\boldsymbol{w}_\star} \right) \left(\frac{1}{n} \sum_{i=1}^n \beta_i \right) + \zeta \left(\frac{1}{n} \sum_{i=1}^n \beta_i^2 y_i \right).
\end{aligned}$$

All the random terms above are zero-mean. Hence,

$$\mathbb{E}[B] = \left((\zeta - \zeta^2) - 2(\zeta^2 - \zeta^3) R_{\boldsymbol{w}_\star} \right) R_{\boldsymbol{q}_\star} . \tag{21}$$

- $\textbf{Term}_{\text{III}} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \zeta\alpha_i - \beta_i\right)\hat{\boldsymbol{\Sigma}}_i\boldsymbol{w}$

Fix any vector $\boldsymbol{v}$:

$$\begin{aligned}
\boldsymbol{v}^T\{\text{Term}_{\text{III}}\} &= \left(\frac{1}{n}\sum_{i=1}^{n} y_i \boldsymbol{v}^T\hat{\boldsymbol{\Sigma}}_i\boldsymbol{w}\right) - \zeta\left(R_{\boldsymbol{q}_\star} + y_i R_{\boldsymbol{w}_\star}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{v}^T\hat{\boldsymbol{\Sigma}}_i\boldsymbol{w}\right) - \left(\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{\gamma}_i^T\boldsymbol{w}\right)\boldsymbol{v}^T\hat{\boldsymbol{\Sigma}}_i\boldsymbol{w}\right)\\
&= \left(1-\zeta R_{\boldsymbol{w}_\star}\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i \boldsymbol{v}^T\hat{\boldsymbol{\Sigma}}_i\boldsymbol{w}\right) - \zeta R_{\boldsymbol{q}_\star}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{v}^T\hat{\boldsymbol{\Sigma}}_i\boldsymbol{w}\right) - \left(\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{\gamma}_i^T\boldsymbol{w}\right)\boldsymbol{v}^T\hat{\boldsymbol{\Sigma}}_i\boldsymbol{w}\right)\\
&= \left(1-\zeta R_{\boldsymbol{w}_\star}\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i \boldsymbol{v}^T\hat{\boldsymbol{\Sigma}}_i\boldsymbol{w}\right) - \zeta R_{\boldsymbol{q}_\star}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{v}^T\hat{\boldsymbol{\Sigma}}_i\boldsymbol{w} - (1-\zeta)\boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{w}\right) - (\zeta-\zeta^2)R_{\boldsymbol{q}_\star}\boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{w}\\
&\quad - \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{\gamma}_i^T\boldsymbol{w}\right)\left(\boldsymbol{v}^T\hat{\boldsymbol{\Sigma}}_i\boldsymbol{w} - (1-\zeta)\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{v}\right) + (1-\zeta)(\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{v})\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\gamma}_i^T\boldsymbol{w})
\end{aligned}$$

From Lemma 6 all random terms above are zero mean. Hence,

$$\mathbb{E}\left[\boldsymbol{v}^T\{\text{Term}_{\text{III}}\}\right] = -(\zeta-\zeta^2)R_{\boldsymbol{q}_\star}\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{v}. \tag{22}$$

Moreover, from Lemma 6 we have the following:

- $\textbf{Term}_{\text{IV}} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \zeta\alpha_i - \beta_i\right)\left(\zeta\alpha_i + \beta_i\right)\boldsymbol{\gamma}_i$

For fixed vector $\boldsymbol{v}$, $\boldsymbol{v}^T\{\text{Term}_{\text{IV}}\} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \zeta\alpha_i - \beta_i\right)\left(\zeta\alpha_i + \beta_i\right)\boldsymbol{v}^T\boldsymbol{\gamma}_i$. Reorganizing, note that

$$\begin{aligned}
\left(y_i - \zeta\alpha_i - \beta_i\right)\left(\zeta\alpha_i + \beta_i\right) &= \zeta y_i(R_{\boldsymbol{q}_\star} + y_i R_{\boldsymbol{w}_\star}) - \zeta^2(R^2_{\boldsymbol{q}_\star} + R^2_{\boldsymbol{w}_\star} + 2y_i R_{\boldsymbol{q}_\star}R_{\boldsymbol{w}_\star}) - 2\zeta R_{\boldsymbol{q}_\star}\beta_i - 2\zeta R_{\boldsymbol{w}_\star}y_i\beta_i + y_i\beta_i - \beta_i^2\\
&= \left(\zeta R_{\boldsymbol{q}_\star} - 2\zeta^2 R_{\boldsymbol{q}_\star}R_{\boldsymbol{w}_\star}\right)y_i + \left(\zeta R_{\boldsymbol{w}_\star} - \zeta^2(R^2_{\boldsymbol{q}_\star} + R^2_{\boldsymbol{w}_\star})\right) - \left(2\zeta R_{\boldsymbol{q}_\star}\right)\beta_i + \left(1 - 2\zeta R_{\boldsymbol{w}_\star}\right)y_i\beta_i - \beta_i^2
\end{aligned}$$

Overall,

$$\begin{aligned}
\boldsymbol{v}^T\{\text{Term}_{\text{IV}}\} &= \left(\zeta R_{\boldsymbol{w}_\star} - \zeta^2(R^2_{\boldsymbol{q}_\star} + R^2_{\boldsymbol{w}_\star})\right)\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{v}^T\boldsymbol{\gamma}_i\right) + \left(\zeta R_{\boldsymbol{q}_\star} - 2\zeta^2 R_{\boldsymbol{q}_\star}R_{\boldsymbol{w}_\star}\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i\left(\boldsymbol{v}^T\boldsymbol{\gamma}_i\right)\right)\\
&\quad + \left(1 - 2\zeta R_{\boldsymbol{w}_\star}\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i\left(\boldsymbol{w}^T\boldsymbol{\gamma}_i\right)\left(\boldsymbol{v}^T\boldsymbol{\gamma}_i\right)\right)\\
&\quad - \left(2\zeta R_{\boldsymbol{q}_\star}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{w}^T\boldsymbol{\gamma}_i\right)\left(\boldsymbol{v}^T\boldsymbol{\gamma}_i\right) - \frac{1-\zeta}{T}\boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{w}\right) - \left(2\zeta R_{\boldsymbol{q}_\star}\right)\frac{1-\zeta}{T}\boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{w}\\
&\quad - \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{v}^T\boldsymbol{\gamma}_i\right)\left(\left(\boldsymbol{w}^T\boldsymbol{\gamma}_i\right)^2 - \frac{(1-\zeta)}{T}\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}\right) + \frac{(1-\zeta)}{T}\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{v}^T\boldsymbol{\gamma}_i\right)\right)
\end{aligned}$$

According to Lemma 6 all random terms above are zero mean. Thus,

$$\mathbb{E}\left[\boldsymbol{v}^T\{\text{Term}_{\text{IV}}\}\right] = -2\zeta R_{\boldsymbol{q}_\star}\frac{1-\zeta}{T}\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{v} \tag{23}$$

- **Combined**

The desired identities (18) and (19) follow by combining all the terms above.

F.2.2 Proof of Lemma 6

First bound: Obvious by boundedness (hence, sub-gaussianity) of y_i and Fact J.1.

Second bound: For convenience set $\widetilde{T} = (1-\zeta)T$ and assume wlog that $\mathcal{R}^c = [\widetilde{T}]$. Recall that

$$\beta_i = \frac{1}{T}\sum_{t=1}^{\widetilde{T}}\boldsymbol{z}_{i,t}^T\boldsymbol{w} = \frac{1-\zeta}{\widetilde{T}}\sum_{t=1}^{\widetilde{T}}\boldsymbol{z}_{i,t}^T\boldsymbol{w}.$$

Also for all t: $\left\|\boldsymbol{z}_{i,t}^T\boldsymbol{w}\right\|_{\psi_2} \le K\|\boldsymbol{w}\|_2$. Thus, from Fact J.1:

$$\|\beta_i\|_{\psi_2} \le \frac{C\sigma(1-\zeta)\|\boldsymbol{w}\|_2}{\sqrt{\widetilde{T}}} = \frac{C\sigma\sqrt{(1-\zeta)}\|\boldsymbol{w}\|_2}{\sqrt{T}} \tag{24}$$

The bound then follows by applying Fact J.1 once more.

For the second term in this bound recall that $y_i \in \{\pm 1\}$ and $\beta_i = \sum_t \boldsymbol{z}_{i,t}^T\boldsymbol{w}/T$. Also, for all $i \in [n]$: $y_i, \{\boldsymbol{z}_{i,t}\}_t$ are zero-mean and independent. Thus (i) $\mathbb{E}[y_i\beta_i] = 0$, and (ii) $\{y_i\boldsymbol{z}_{i,t} \overset{D}{\sim} \boldsymbol{z}_{i,t}$ and $y_i\boldsymbol{z}_{i,t} \perp y_i\boldsymbol{z}_{i,t'}\} \implies y_i\beta_i \overset{D}{\sim} \beta_i$. Thus, the same bound as the first term holds.

Third bound: It is easy to compute

$$\mathbb{E}[\left(\boldsymbol{\gamma}_i^T\boldsymbol{w}\right)\left(\boldsymbol{\gamma}_i^T\boldsymbol{v}\right)] = \frac{1}{T^2}\sum_{t=1}^{\widetilde{T}}\sum_{t'=1}^{\widetilde{T}}\mathbb{E}[\boldsymbol{w}^T\boldsymbol{z}_{i,t}\boldsymbol{z}_{i,t'}\boldsymbol{v}] = \frac{1-\zeta}{T}\boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{w}, \tag{25}$$

and, using (24)

$$\left\|\left(\boldsymbol{\gamma}_i^T\boldsymbol{w}\right)\left(\boldsymbol{\gamma}_i^T\boldsymbol{v}\right) - \mathbb{E}[\left(\boldsymbol{\gamma}_i^T\boldsymbol{w}\right)\left(\boldsymbol{\gamma}_i^T\boldsymbol{v}\right)]\right\|_{\psi_1} \le C\left\|\left(\boldsymbol{\gamma}_i^T\boldsymbol{w}\right)\left(\boldsymbol{\gamma}_i^T\boldsymbol{v}\right)\right\|_{\psi_1} \le C\left\|\boldsymbol{\gamma}_i^T\boldsymbol{w}\right\|_{\psi_2}\left\|\boldsymbol{\gamma}_i^T\boldsymbol{v}\right\|_{\psi_2} = \frac{C\sigma^2(1-\zeta)\|\boldsymbol{w}\|_2\|\boldsymbol{v}\|_2}{T}. \tag{26}$$

Since $\left(\boldsymbol{\gamma}_i^T\boldsymbol{w}\right)\left(\boldsymbol{\gamma}_i^T\boldsymbol{v}\right), i \in [n]$ are independent, the desired bound on the first term follows from Fact J.1.

Consider now the second term. By independence of $y_i, \boldsymbol{\gamma}_i$ it holds that $\mathbb{E}[y_i\left(\boldsymbol{\gamma}_i^T\boldsymbol{w}\right)\left(\boldsymbol{\gamma}_i^T\boldsymbol{v}\right)] = 0$. Arguing as we did above for the second bound, $y_i\boldsymbol{\gamma}_i^T\boldsymbol{w} \overset{D}{\sim} \boldsymbol{\gamma}_i^T\boldsymbol{w}$. Hence, the subexponential bound is the same as for the first term.

Fourth bound: First, it is easy to compute that for all $i \in [n]$:

$$\mathbb{E}[\boldsymbol{w}^T\boldsymbol{\Sigma}_i\boldsymbol{v}] = \frac{1}{T}\mathbb{E}[\boldsymbol{w}^T\boldsymbol{Z}_i^T\boldsymbol{Z}_i\boldsymbol{v}] = \frac{1}{T}\sum_{t=1}^{\widetilde{T}}\mathbb{E}[\boldsymbol{w}^T\boldsymbol{z}_{i,t}^T\boldsymbol{z}_{i,t}\boldsymbol{v}] = \frac{\widetilde{T}}{T}\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{v} = (1-\zeta)\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{v}.$$

Thus,

$$\boldsymbol{w}^T\hat{\boldsymbol{\Sigma}}_i\boldsymbol{v} - \mathbb{E}[\boldsymbol{w}^T\hat{\boldsymbol{\Sigma}}_i\boldsymbol{v}] = \boldsymbol{w}^T\hat{\boldsymbol{\Sigma}}_i\boldsymbol{v} - (1-\zeta)\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{v} = \frac{1}{T}\sum_{t=1}^{\widetilde{T}}\left((\boldsymbol{z}_{i,t}^T\boldsymbol{w})(\boldsymbol{z}_{i,t}^T\boldsymbol{v}) - \boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{v}\right)$$

and so

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{w}^T\hat{\boldsymbol{\Sigma}}_i\boldsymbol{v} - (1-\zeta)\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{v} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{\widetilde{T}}\left((\boldsymbol{z}_{i,t}^T\boldsymbol{w})(\boldsymbol{z}_{i,t}^T\boldsymbol{v}) - \boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{v}\right).$$

Now, each random variable in the double sum above is independent and such that

$$\|(\boldsymbol{z}_{i,t}^T\boldsymbol{w})(\boldsymbol{z}_{i,t}^T\boldsymbol{v}) - \boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{v}\|_{\psi_1} \le 2\|(\boldsymbol{z}_{i,t}^T\boldsymbol{w})(\boldsymbol{z}_{i,t}^T\boldsymbol{v})\|_{\psi_1} \le 2\|\boldsymbol{z}_{i,t}^T\boldsymbol{w}\|_{\psi_2}\|\boldsymbol{z}_{i,t}^T\boldsymbol{v}\|_{\psi_2} \le C\sigma^2\|\boldsymbol{w}\|_2\|\boldsymbol{v}\|_2. \tag{27}$$

Hence, from Fact J.1,

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{w}^T\hat{\boldsymbol{\Sigma}}_i\boldsymbol{v} - (1-\zeta)\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{v}\right\|_{\psi_1} \le \frac{C\sigma^2\sqrt{1-\zeta}\|\boldsymbol{w}\|_2\|\boldsymbol{v}\|_2}{\sqrt{nT}}$$

The bound for the second term follows along the same lines. The two key observations are that (i) $\mathbb{E}[y_i\boldsymbol{w}^T\hat{\boldsymbol{\Sigma}}_i\boldsymbol{v}] = 0$ because y_i and $\boldsymbol{z}_{i,t}$ are independent, and, (ii)

$$y_i\boldsymbol{w}^T\hat{\boldsymbol{\Sigma}}_i\boldsymbol{v} = \frac{1}{T}\boldsymbol{w}^Ty_i\boldsymbol{Z}_i^T\boldsymbol{Z}_i\boldsymbol{v} \overset{D}{\sim} \frac{1}{T}\boldsymbol{w}^T\widetilde{\boldsymbol{Z}}_i^T\boldsymbol{Z}_i\boldsymbol{v} = \frac{1}{T}\sum_{t=1}^{\widetilde{T}}(\widetilde{\boldsymbol{z}}_{i,t}^T\boldsymbol{w})(\boldsymbol{z}_{i,t}^T\boldsymbol{v})$$

where $\widetilde{\boldsymbol{Z}}_i$ is an independent copy of $\boldsymbol{Z}_i$.

Fifth bound:From (28) and (25), we have for all $i \in [n]$ that

$$\left\| \left(\boldsymbol{\gamma}_i^T \boldsymbol{w}\right)\left(\boldsymbol{\gamma}_i^T \boldsymbol{v}\right) - \frac{1-\zeta}{T}\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w} \right\|_{\psi_1} \le \frac{C\sigma^2(1-\zeta)\|\boldsymbol{w}\|_2^2}{T}.$$

Moreover, recall from Eq. (24) that

$$\left\|\boldsymbol{\gamma}_i^T \boldsymbol{v}\right\|_{\psi_2} \le \frac{C\sigma\sqrt{1-\zeta}\|\boldsymbol{v}\|_2}{\sqrt{T}}.$$

Combining the above two displays and applying Fact J.2 we find for all $i \in [n]$ that

$$\left\| \left(\left(\boldsymbol{\gamma}_i^T \boldsymbol{w}\right)^2 - \frac{1-\zeta}{T}\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w} \right)\boldsymbol{\gamma}_i^T \boldsymbol{v} \right\|_{\psi_{2/3}} \le \frac{C\sigma^3(1-\zeta)^{3/2}\|\boldsymbol{w}\|_2^2\|\boldsymbol{v}\|_2}{T^{3/2}}. \tag{28}$$

The desired bound follows from the above after using Fact J.3.

Sixth bound: From Eq. (27):

$$\left\|\boldsymbol{w}^T\hat{\boldsymbol{\Sigma}}_i\boldsymbol{v} - (1-\zeta)\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{v}\right\|_{\psi_1} \le \frac{C\sigma^2\sqrt{1-\zeta}\|\boldsymbol{w}\|_2\|\boldsymbol{v}\|_2}{\sqrt{T}}$$

and from Eq. (24)

$$\left\|\boldsymbol{\gamma}_i^T \boldsymbol{w}\right\|_{\psi_2} \le \frac{C\sigma\sqrt{1-\zeta}\|\boldsymbol{w}\|_2}{\sqrt{T}}.$$

Next we use Fact J.2 with $\alpha = 2$ and $\beta = 1$ to find that

$$\left\|\left(\boldsymbol{\gamma}_i^T \boldsymbol{w}\right)\left(\boldsymbol{w}^T\hat{\boldsymbol{\Sigma}}_i\boldsymbol{v} - (1-\zeta)\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{v}\right)\right\|_{\psi_{2/3}} \le \frac{CK^3(1-\zeta)\|\boldsymbol{w}\|_2^2\|\boldsymbol{v}\|_2}{T}$$

Next we use Fact J.3 which allows us to conclude that

$$\left\|\frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{\gamma}_i^T \boldsymbol{w}\right)\left(\boldsymbol{w}^T\hat{\boldsymbol{\Sigma}}_i\boldsymbol{v} - (1-\zeta)\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{v}\right)\right\|_{\psi_{2/3}} \le \frac{CK^3(1-\zeta)\|\boldsymbol{w}\|_2^2\|\boldsymbol{v}\|_2}{T\sqrt{n}}\log n$$

Finally, the zero-mean property follows since

$$\begin{aligned}
\mathbb{E}[\left(\mathbb{1}^T\boldsymbol{Z}_i\boldsymbol{w}\right)\left(\boldsymbol{w}^T\boldsymbol{Z}_i\boldsymbol{Z}_i^T\boldsymbol{v}\right)] &= \sum_{t=1}^{\widetilde{T}}\sum_{t'=1}^{\widetilde{T}}\mathbb{E}[\left(\boldsymbol{z}_{i,t}^T\boldsymbol{w}\right)\left(\boldsymbol{w}^T\boldsymbol{z}_{i,t'}\boldsymbol{z}_{i,t'}^T\boldsymbol{v}\right)] = \sum_{t=1}^{\widetilde{T}}\sum_{t'=1}^{\widetilde{T}}\mathbb{E}[\boldsymbol{w}^T\boldsymbol{z}_{i,t}\operatorname{tr}\left(\boldsymbol{z}_{i,t'}\boldsymbol{z}_{i,t'}^T\boldsymbol{v}\boldsymbol{w}^T\right)] \\
&= \widetilde{T}^2\mathbb{E}[\operatorname{tr}\left(\boldsymbol{z}^T\boldsymbol{w}\right)\operatorname{tr}\left(\boldsymbol{z}\boldsymbol{z}^T\boldsymbol{v}\boldsymbol{w}^T\right)] = \widetilde{T}^2\mathbb{E}[\operatorname{tr}\left((\boldsymbol{z}^T\boldsymbol{w})\otimes(\boldsymbol{z}\boldsymbol{z}^T\boldsymbol{v}\boldsymbol{w}^T\right))] \\
&= \widetilde{T}^2\operatorname{tr}\left(\mathbb{E}[(\boldsymbol{z}^T\otimes\boldsymbol{z}\boldsymbol{z}^T)](\boldsymbol{w}\otimes\boldsymbol{v}\boldsymbol{w}^T)\right) = 0
\end{aligned}$$

where the last equality follows by the zero third moment property in Assumption 1.a.

F.3 Finite-sample analysis: First and second gradient steps

The lemma below studies the deviation of the first-step of GD $\widehat{\boldsymbol{w}}_1$ with respect to its population counterpart $\boldsymbol{w}_1$. Provided that n and $n\zeta T/d$ are larger than appropriate functions of other problem parameters, then the deviations are of small multiplicative order. We remark that the constants below have not been optimized (eg. the factor $1/2$ in (30) is arbitrary and can be replaced by and small positive constant by appropriately increasing the constant C in (29)).

Lemma 7 (First gradient step)**.** *Consider the one-step population and finite updates $\boldsymbol{w}_1 = \eta\boldsymbol{G}_{\boldsymbol{w}}(0,0)$ and $\widehat{\boldsymbol{w}}_1 = \eta\widehat{\boldsymbol{G}}_{\boldsymbol{w}}(0,0)$, respectively. For convenience denote $R_{\boldsymbol{w}_\star} = \boldsymbol{w}_1^T\boldsymbol{w}_\star, R_{\boldsymbol{q}_\star} = \boldsymbol{w}_1^T\boldsymbol{q}_\star$ and $\widehat{R}_{\boldsymbol{w}_\star} = \widehat{\boldsymbol{w}}_1^T\boldsymbol{w}_\star, \widehat{R}_{\boldsymbol{q}_\star} = \widehat{\boldsymbol{w}}_1^T\boldsymbol{q}_\star$ their finite-sample counterparts. Suppose for large enough absolute constant $C > 0$ and any $u > 0$,*

$$\sqrt{n} \ge CuQ/W \qquad \text{and} \qquad \sqrt{n\zeta T} \ge Cu\frac{\sigma}{W}\sqrt{\zeta^{-1}-1} \tag{29}$$

Then, for absolute constants $c, c' > 0$ with probability at least $1 - c'e^{-cu^2}$

$$\left|\widehat{R}_{\boldsymbol{w}_\star} - R_{\boldsymbol{w}_\star}\right| \le R_{\boldsymbol{w}_\star}/2 \qquad \text{and} \qquad \left|\widehat{R}_{\boldsymbol{q}_\star} - R_{\boldsymbol{q}_\star}\right| \le \eta\zeta QW/2 \tag{30}$$

Additionally, if

$$\sqrt{n} \geq CuQ/W \qquad \text{and} \qquad \sqrt{n\zeta T} \geq Cu\frac{\sigma}{W}\sqrt{\zeta^{-1}-1}\sqrt{d} \tag{31}$$

then with the same probability

$$\left|\|\widehat{\boldsymbol{w}}_1\| - \|\boldsymbol{w}_1\|\right| \leq \|\boldsymbol{w}_1\|/2. \tag{32}$$

Proof. Note that the conclusions of the lemma are all homogeneous in η. Hence, without loss of generality, set $\eta = 1$.

By Lemma 3,

$$\widehat{\boldsymbol{w}}_1 = \widehat{\boldsymbol{G}}_{\boldsymbol{w}}(0,0) = \zeta \boldsymbol{w}_\star + \zeta \boldsymbol{q}_\star \left(\frac{1}{n}\sum_{i\in[n]} y_i\right) + \frac{1}{n}\sum_i y_i \boldsymbol{\gamma}_i. \tag{33}$$

and recall $\boldsymbol{w}_1 = \boldsymbol{G}_{\boldsymbol{w}}(0,0) = \zeta \boldsymbol{w}_\star$ (thus, $R_{\boldsymbol{w}_\star} = \zeta W^2$). From these, and also using Lemma 6 , for any $u > 0$ with probability at least $1 - 2e^{-cu^2}$

$$\left|\widehat{R}_{\boldsymbol{w}_\star} - R_{\boldsymbol{w}_\star}\right| = \left|\boldsymbol{w}_\star^T(\widehat{\boldsymbol{w}}_1 - \boldsymbol{w}_1)\right| \leq \zeta\,|\rho|\,WQ\left|\frac{1}{n}\sum_{i\in[n]} y_i\right| + \left|\frac{1}{n}\sum_i y_i \boldsymbol{\gamma}_i^T \boldsymbol{w}_\star\right| \leq \frac{Cu\zeta\,|\rho|\,WQ}{\sqrt{n}} + \frac{Cu\sigma\sqrt{1-\zeta}\sqrt{\zeta}W}{\sqrt{n\zeta T}}$$
$$\leq \frac{\zeta W^2}{2} = \frac{R_{\boldsymbol{w}_\star}}{2}.$$

where the last inequality follows by assuming $n, \zeta T$ large enough as in the condition of the lemma. Similarly,

$$\left|\widehat{R}_{\boldsymbol{q}_\star} - R_{\boldsymbol{q}_\star}\right| = \left|\boldsymbol{q}_\star^T(\widehat{\boldsymbol{w}}_1 - \boldsymbol{w}_1)\right| \leq \zeta Q^2\left|\frac{1}{n}\sum_{i\in[n]} y_i\right| + \left|\frac{1}{n}\sum_i y_i \boldsymbol{\gamma}_i^T \boldsymbol{q}_\star\right| \leq \frac{Cu\zeta Q^2}{\sqrt{n}} + \frac{Cu\sigma\sqrt{1-\zeta}\sqrt{\zeta}W}{\sqrt{n\zeta T}} \leq \frac{\zeta QW}{2}$$

for sufficiently large n as in (29).

Finally,

$$\|\widehat{\boldsymbol{w}}_1\| - \|\boldsymbol{w}_1\| \leq \|\widehat{\boldsymbol{w}}_1 - \boldsymbol{w}_1\| \leq \frac{Cu\zeta Q}{\sqrt{n}} + \frac{Cu\sigma\sqrt{1-\zeta}\sqrt{\zeta}\sqrt{d}}{\sqrt{n\zeta T}} \leq \frac{\zeta W}{2} = \frac{\|\boldsymbol{w}_1\|}{2} \tag{34}$$

where, again, the last inequality follows by assuming $n, \zeta T$ large enough as stated in the lemma. In the second inequality, we used from Lemma 6 that $\sum_{i\in[n]} y_i\boldsymbol{\gamma}_i/n$ is $C\sigma\sqrt{1-\zeta}/\sqrt{nT}$-subGaussian and applied Fact J.4 to get a high-probability bound on its euclidean norm. □

Next, we move on to the second gradient update in the direction of $\widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0, \widehat{\boldsymbol{w}}_1)$. Recall our goal of controlling the relevance scores of signal and noisy tokens. The first lemma below takes a step in this direction by computing the signal and noise relevance scores assuming access to the population gradient $\boldsymbol{G}_{\boldsymbol{q}}(0, \widehat{\boldsymbol{w}}_1) := \mathbb{E}\left[\widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0, \widehat{\boldsymbol{w}}_1)\right]$.

Lemma 8 ($\widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0,\cdot)$ control: Expectation term)**.** *Let $\boldsymbol{G}_{\boldsymbol{q}}(0, \boldsymbol{w}_1) = \mathbb{E}\left[\widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0, \boldsymbol{w}_1)\right]$ be the expectation of a step in the $\boldsymbol{q}$-gradient evaluated at $(0, \widehat{\boldsymbol{w}}_1)$ and recall that $\widehat{\boldsymbol{w}}_1 = \eta\widehat{\boldsymbol{G}}_{\boldsymbol{w}}(0,0)$ for $\eta > 0$. Suppose $\widehat{\boldsymbol{w}}_1$ satisfies (30) and (32). Further assume that the step-size η satisfies the following for sufficiently small absolute constant $c_\eta > 0$:*

$$\eta < c_\eta\left(W^{-2} \vee Q^{-2} \vee \sigma^{-2}\right) \tag{35}$$

Then, for $y \in \{\pm 1\}$ it holds that

$$(\boldsymbol{q}_\star + y\boldsymbol{w}_\star)^T \boldsymbol{G}_{\boldsymbol{q}}(0, \hat{\boldsymbol{w}}_1) \geq \eta\zeta(\zeta - \zeta^2)W^2Q\left(\left(15 - 8\rho^2 - 6\,|\rho|\right)Q - (210\,|\rho| + 33)\,W\right)/64 \tag{36}$$

Moreover, for fresh noise variables $\boldsymbol{z}_t, t \in \mathcal{R}^c$ satisfying Assumption 1.a, it holds that

$$\boldsymbol{z}_t^T \boldsymbol{G}_{\boldsymbol{q}}(0, \hat{\boldsymbol{w}}_1) \overset{D}{=} \eta\zeta\left(\zeta - \zeta^2\right)W^2Q\,\sigma\sqrt{3}\sqrt{(3/2)^4 + 2\left((1 + 1/(4\cdot 16^2))\right)\rho^2 + 1/4)}\cdot G_t, \quad G_t \sim \mathcal{SN}(1) \tag{37}$$

Proof. Fix any $\boldsymbol{v}$ and recall the notation of Lemma 7. With these, we have from Lemma 5 that

$$\begin{aligned}
\boldsymbol{v}^T \boldsymbol{G}_{\boldsymbol{q}}(0,\widehat{\boldsymbol{w}}_1) &= \left((\zeta-\zeta^2)\left(\widehat{R}_{\boldsymbol{w}_\star}+\sigma^2\|\widehat{\boldsymbol{w}}_1\|^2/T\right)-(\zeta^2-\zeta^3)\left(\widehat{R}^2_{\boldsymbol{w}_\star}+\widehat{R}^2_{\boldsymbol{q}_\star}\right)\right)\boldsymbol{v}^T\boldsymbol{q}_\star \\
&\quad + \left((\zeta-\zeta^2)-2(\zeta^2-\zeta^3)\widehat{R}_{\boldsymbol{w}_\star}\right)\widehat{R}_{\boldsymbol{q}_\star}\boldsymbol{v}^T\boldsymbol{w}_\star - \sigma^2(1+2/T)(\zeta-\zeta^2)\widehat{R}_{\boldsymbol{q}_\star}\boldsymbol{v}^T\widehat{\boldsymbol{w}}_1 \\
&= (\zeta-\zeta^2)\widehat{R}_{\boldsymbol{w}_\star}\left(1+\sigma^2\|\widehat{\boldsymbol{w}}_1\|^2/(T\widehat{R}_{\boldsymbol{w}_\star})-\zeta\left(\widehat{R}_{\boldsymbol{w}_\star}+\widehat{R}^2_{\boldsymbol{q}_\star}/\widehat{R}_{\boldsymbol{w}_\star}\right)\right)\boldsymbol{v}^T\boldsymbol{q}_\star \\
&\quad + (\zeta-\zeta^2)\widehat{R}_{\boldsymbol{q}_\star}\left(1-2\zeta\widehat{R}_{\boldsymbol{w}_\star}\right)\boldsymbol{v}^T\boldsymbol{w}_\star - \sigma^2(1+2/T)(\zeta-\zeta^2)\widehat{R}_{\boldsymbol{q}_\star}\boldsymbol{v}^T\widehat{\boldsymbol{w}}_1 \\
&= (\zeta-\zeta^2)\widehat{R}_{\boldsymbol{w}_\star}\underbrace{\left(1+\sigma^2\|\widehat{\boldsymbol{w}}_1\|^2/(T\widehat{R}_{\boldsymbol{w}_\star})-\zeta\left(\widehat{R}_{\boldsymbol{w}_\star}+\widehat{R}^2_{\boldsymbol{q}_\star}/\widehat{R}_{\boldsymbol{w}_\star}\right)\right)}_{:=\widehat{C}_1}\boldsymbol{v}^T\boldsymbol{q}_\star \\
&\quad + (\zeta-\zeta^2)\widehat{R}_{\boldsymbol{q}_\star}\underbrace{\left(1-2\zeta\widehat{R}_{\boldsymbol{w}_\star}-\eta\zeta\sigma^2(1+2/T)\right)}_{:=\widehat{C}_2}\boldsymbol{v}^T\boldsymbol{w}_\star - \underbrace{\eta\sigma^2(1+2/T)}_{:=\widehat{C}_3}(\zeta-\zeta^2)\widehat{R}_{\boldsymbol{q}_\star}\boldsymbol{v}^T\boldsymbol{\delta}_1
\end{aligned}$$

where we used (33) and set $\boldsymbol{\delta}_1 := (\widehat{\boldsymbol{w}}_1-\boldsymbol{w}_1)/\eta = \zeta\boldsymbol{q}_\star\left(\frac{1}{n}\sum_{i\in[n]}y_i\right)+\frac{1}{n}\sum_i y_i\boldsymbol{\gamma}_i$.

Recall that $R_{\boldsymbol{w}_\star}/2 \le \widehat{R}_{\boldsymbol{w}_\star} \le 3R_{\boldsymbol{w}_\star}/2$, $R_{\boldsymbol{q}_\star}-\eta\zeta QW/2 \le \widehat{R}_{\boldsymbol{q}_\star} \le R_{\boldsymbol{q}_\star}+\eta\zeta QW/2$ and $\|\boldsymbol{w}_1\|/2 \le \|\widehat{\boldsymbol{w}}_1\| \le 3\|\boldsymbol{w}_1\|/2$. Also, $R_{\boldsymbol{w}_\star}=\eta\zeta W^2$ $R_{\boldsymbol{q}_\star}=\eta\zeta\rho WQ$ and $\|\boldsymbol{w}_1\|=\eta\zeta W$. With these, we can set step size η small enough such that $\widehat{C}_1 \in [1/2,3/2]$, $\widehat{C}_2 \in [-1/8,1]$ and $\widehat{C}_3 \in [0,1/(16\zeta)]$. Also, recall from (34) $\|\boldsymbol{\delta}_1\| \le \zeta W/2$; thus,

$$\widehat{C}_3\left|\widehat{R}_{\boldsymbol{q}_\star}\right|\|\boldsymbol{\delta}_1\| \le \frac{1}{16\zeta}\eta\zeta QW(|\rho|+1/2)\frac{\zeta W}{2} = \eta\zeta\frac{(|\rho|+1/2)QW^2}{32}.$$

With these, we arrive at the following:

$$\boldsymbol{q}_\star^T\boldsymbol{G}_{\boldsymbol{q}}(0,\hat{\boldsymbol{w}}_1) \ge \eta\zeta(\zeta-\zeta^2)\left(W^2Q^2/4-|\rho|(|\rho|+1/2)Q^2W^2/8-(|\rho|+1/2)Q^2W^2/32\right) \ge \eta\zeta(\zeta-\zeta^2)W^2Q^2\left(15-8\rho^2-6|\rho|\right)$$

and for $y\in\{\pm1\}$

$$\left|y\boldsymbol{w}_\star^T\boldsymbol{G}_{\boldsymbol{q}}(0,\hat{\boldsymbol{w}}_1)\right| \le \eta\zeta\left(\zeta-\zeta^2\right)W^3Q\left(|\rho|(9/4+1+1/32)+(1/2+1/64)\right) = \eta\zeta\left(\zeta-\zeta^2\right)W^3Q\left(|\rho|(105/32)+33/64\right).$$

The above two displays put together yield (36).

To see (37), note that $\boldsymbol{z}_t^T\boldsymbol{w}_\star \sim \mathcal{SN}(\sigma W)$, $\boldsymbol{z}_t^T\boldsymbol{q}_\star \sim \mathcal{SN}(\sigma Q)$ and $\boldsymbol{z}_t^T\boldsymbol{\delta}_1 \sim \mathcal{SN}(\sigma\|\boldsymbol{\delta}_1\|)$. Thus, the noise relevant scores are IID subgaussian random variables with sub-gaussian constant $\widehat{\sigma}$ upper bounded as follows

$$\boldsymbol{z}_t^T\boldsymbol{G}_{\boldsymbol{q}}(0,\hat{\boldsymbol{w}}_1) = \eta\zeta\left(\zeta-\zeta^2\right)W^2Q\,\widehat{\sigma}\cdot G_t,\ G_t\sim\mathcal{SN}(1) \quad\text{and}\quad \widehat{\sigma} \le \sigma\sqrt{3}\sqrt{(3/2)^4+2\left((1+1/(4\cdot16^2))\rho^2+1/4\right)}. \tag{38}$$

□

The next lemma controls the effect on the relevance scores of the deviation term $\widetilde{\boldsymbol{G}}_{\boldsymbol{q}}(0,\widehat{\boldsymbol{w}}) = \widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0,\widehat{\boldsymbol{w}}) - \boldsymbol{G}_{\boldsymbol{q}}(0,\widehat{\boldsymbol{w}})$.

Lemma 9 ($\widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0,\cdot)$ control: Deviation term)**.** *Let $\widetilde{\boldsymbol{G}}_{\boldsymbol{q}}(0,\widehat{\boldsymbol{w}}_1) := \widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0,\widehat{\boldsymbol{w}}_1) - \boldsymbol{G}_{\boldsymbol{q}}(0,\widehat{\boldsymbol{w}}_1)$ and suppose $\widehat{\boldsymbol{w}}_1$ satisfies (30) and (32). Fix any $u>0$ and any small constant $c_1>0$. Then, there exists small enough constant c_η (dependent on c_1) such that if step-size η is small enough as per (35) the following statements hold.*

First, for signal tokens, there exist positive constants C, c', c such that with probability at least $1-c'e^{-cu^{2/3}}$

$$\left|(\boldsymbol{q}_\star+y\boldsymbol{w}_\star)^T\widetilde{\boldsymbol{G}}_{\boldsymbol{q}}(0,\widehat{\boldsymbol{w}}_1)\right| \le uCc_1(W^2\vee Q^2)\left(1\vee\frac{\sigma\vee\sigma^2}{\sqrt{T}}\vee\frac{\sigma^3}{T}\right)\frac{\log(n)}{\sqrt{n}}, \quad y\in\{\pm1\} \tag{39}$$

Second, for fresh noise variables $\boldsymbol{z}_t, t\in\mathcal{R}^c$ satisfying Assumption 1.a, it holds that

$$\boldsymbol{z}_t^T\boldsymbol{G}_{\boldsymbol{q}}(0,\hat{\boldsymbol{w}}_1)\sim\mathcal{SN}(\widetilde{\sigma}) \quad \textit{where w.p.} \ge 1-c'de^{-cu^{2/3}} \quad \widetilde{\sigma} \le u\sigma Cc_1\left(W\vee Q\vee\frac{\sigma\vee\sigma^2}{\sqrt{T}}\vee\frac{\sigma^3}{T}\right)\frac{\log(n)}{\sqrt{n}}. \tag{40}$$

Proof. We study each one of the terms of $\widetilde{\boldsymbol{G}}_{\boldsymbol{q}}(0,\widehat{\boldsymbol{w}}_1)$ in (19) separately. We repeat the terms here for convenience, also noting the substitutions $R_{\boldsymbol{w}_\star} \leftarrow \widehat{R}_{\boldsymbol{w}_\star} = \widehat{\boldsymbol{w}}_1^T \boldsymbol{w}_\star$ and $R_{\boldsymbol{q}_\star} \leftarrow \widehat{R}_{\boldsymbol{q}_\star} = \widehat{\boldsymbol{w}}_1^T \boldsymbol{q}_\star$.

$$
\begin{aligned}
\boldsymbol{v}^T \widetilde{\boldsymbol{G}}_{\boldsymbol{q}}(0,\widehat{\boldsymbol{w}}_1) = & \Big[\big((\zeta-\zeta^2) - 2(\zeta^2-\zeta^3)\widehat{R}_{\boldsymbol{w}_\star}\big)\,\widehat{R}_{\boldsymbol{q}_\star} R_{\boldsymbol{v},\boldsymbol{q}_\star} + \big((\zeta-\zeta^2)\widehat{R}_{\boldsymbol{w}_\star} - (\zeta^2-\zeta^3)\big(\widehat{R}^2_{\boldsymbol{w}_\star} + \widehat{R}^2_{\boldsymbol{q}_\star}\big)\big) R_{\boldsymbol{v},\boldsymbol{w}_\star}\Big]\left(\frac{1}{n}\sum_{i=1}^n y_i\right) \\
& + \Big[\big(-\zeta+2\zeta^2\big)\,\widehat{R}_{\boldsymbol{q}_\star} R_{\boldsymbol{v},\boldsymbol{q}_\star} + \big(-\zeta + (-\zeta+2\zeta^2)\widehat{R}_{\boldsymbol{w}_\star}\big) R_{\boldsymbol{v},\boldsymbol{w}_\star} + (1-\zeta)\widehat{\boldsymbol{w}}_1^T\boldsymbol{\Sigma}\boldsymbol{v}\Big]\left(\frac{1}{n}\sum_{i=1}^n \big(\boldsymbol{\gamma}_i^T\widehat{\boldsymbol{w}}_1\big)\right) \\
& \quad + \left[\zeta\widehat{R}_{\boldsymbol{w}_\star} - \zeta^2(\widehat{R}^2_{\boldsymbol{q}_\star} + \widehat{R}^2_{\boldsymbol{w}_\star}) + \frac{(1-\zeta)}{T}\widehat{\boldsymbol{w}}_1^T\boldsymbol{\Sigma}\widehat{\boldsymbol{w}}_1\right]\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{v}^T\boldsymbol{\gamma}_i\right) \\
& \quad + \Big[\big(-\zeta + (-\zeta+2\zeta^2)\widehat{R}_{\boldsymbol{w}_\star}\big)\,R_{\boldsymbol{v},\boldsymbol{q}_\star} + \big(-\zeta+2\zeta^2\big)\,\widehat{R}_{\boldsymbol{q}_\star} R_{\boldsymbol{v},\boldsymbol{w}_\star}\Big]\left(\frac{1}{n}\sum_{i=1}^n y_i\big(\boldsymbol{\gamma}_i^T\widehat{\boldsymbol{w}}_1\big)\right) \\
& \quad + \big[\zeta\widehat{R}_{\boldsymbol{q}_\star} - 2\zeta^2\widehat{R}_{\boldsymbol{q}_\star}\widehat{R}_{\boldsymbol{w}_\star}\big]\left(\frac{1}{n}\sum_{i=1}^n y_i\big(\boldsymbol{v}^T\boldsymbol{\gamma}_i\big)\right) \\
& + \zeta R_{\boldsymbol{v},\boldsymbol{q}_\star}\left(\frac{1}{n}\sum_{i=1}^n \big(\boldsymbol{\gamma}_i^T\widehat{\boldsymbol{w}}_1\big)^2 - \frac{(1-\zeta)}{T}\widehat{\boldsymbol{w}}_1^T\boldsymbol{\Sigma}\widehat{\boldsymbol{w}}_1\right) \\
& \quad + \zeta R_{\boldsymbol{v},\boldsymbol{w}_\star}\left(\frac{1}{n}\sum_{i=1}^n \big(\boldsymbol{\gamma}_i^T\widehat{\boldsymbol{w}}_1\big)^2 y_i\right) \\
& \quad + \big[1-2\zeta\widehat{R}_{\boldsymbol{w}_\star}\big]\left(\frac{1}{n}\sum_{i=1}^n y_i\big(\widehat{\boldsymbol{w}}_1^T\boldsymbol{\gamma}_i\big)\big(\boldsymbol{v}^T\boldsymbol{\gamma}_i\big)\right) \\
& + \big(1-\zeta\widehat{R}_{\boldsymbol{w}_\star}\big)\left(\frac{1}{n}\sum_{i=1}^n y_i\boldsymbol{v}^T\hat{\boldsymbol{\Sigma}}_i\widehat{\boldsymbol{w}}_1\right) \\
& \quad - \zeta\widehat{R}_{\boldsymbol{q}_\star}\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{v}^T\hat{\boldsymbol{\Sigma}}_i\widehat{\boldsymbol{w}}_1 - (1-\zeta)\boldsymbol{v}^T\boldsymbol{\Sigma}\widehat{\boldsymbol{w}}_1\right) \\
& - \big[2\zeta\widehat{R}_{\boldsymbol{q}_\star}\big]\left(\frac{1}{n}\sum_{i=1}^n \big(\widehat{\boldsymbol{w}}_1^T\boldsymbol{\gamma}_i\big)\big(\boldsymbol{v}^T\boldsymbol{\gamma}_i\big) - \frac{1-\zeta}{T}\boldsymbol{v}^T\boldsymbol{\Sigma}\widehat{\boldsymbol{w}}_1\right) \\
& - \left(\frac{1}{n}\sum_{i=1}^n \big(\boldsymbol{v}^T\boldsymbol{\gamma}_i\big)\left(\big(\widehat{\boldsymbol{w}}_1^T\boldsymbol{\gamma}_i\big)^2 - \frac{(1-\zeta)}{T}\widehat{\boldsymbol{w}}_1^T\boldsymbol{\Sigma}\widehat{\boldsymbol{w}}_1\right)\right) \\
& - \left(\frac{1}{n}\sum_{i=1}^n \big(\boldsymbol{\gamma}_i^T\widehat{\boldsymbol{w}}_1\big)\big(\boldsymbol{v}^T\hat{\boldsymbol{\Sigma}}_i\widehat{\boldsymbol{w}}_1 - (1-\zeta)\widehat{\boldsymbol{w}}_1^T\boldsymbol{\Sigma}\boldsymbol{v}\big)\right). \qquad (41)
\end{aligned}
$$

Recall from the lemma assumption that (30) holds and from $R_{\boldsymbol{w}_\star} = \eta\zeta W^2$ and $R_{\boldsymbol{q}_\star} = \eta\zeta\rho WQ$, that $1/2\eta\zeta W^2 \le \widehat{R}_{\boldsymbol{w}_\star} \le 3/2\eta\zeta W^2$ and $\eta\zeta(\rho-1/2)WQ \le \big|\widehat{R}_{\boldsymbol{q}_\star}\big| \le \eta\zeta(\rho+1/2)WQ$. The observation is that we can choose step-size η small enough (as stated in (35)) to bound (in absolute value) all the coefficients in (41) (aka all terms in square brackets) that include $\widehat{R}_{\boldsymbol{w}_\star}, \widehat{R}_{\boldsymbol{q}_\star}$. Therefore, for any small positive constant $c_1 > 0$ it can be checked that there is sufficiently small constant c_η that determines

step-size η in (35) such that

$$
\begin{aligned}
\left|\boldsymbol{v}^T \widetilde{\boldsymbol{G}}_{\boldsymbol{q}}(0, \widehat{\boldsymbol{w}}_1)\right| \leq\; & c_1 \left[|R_{\boldsymbol{v},\boldsymbol{q}_\star}| + |R_{\boldsymbol{v},\boldsymbol{w}_\star}|\right] \left|\frac{1}{n}\sum_{i=1}^n y_i\right| \\
& + \left[c_1 |R_{\boldsymbol{v},\boldsymbol{q}_\star}| + c_1 |R_{\boldsymbol{v},\boldsymbol{w}_\star}| + (1-\zeta)\sigma^2 \left|\widehat{\boldsymbol{w}}_1^T \boldsymbol{v}\right|\right] \left|\frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{\gamma}_i^T \widehat{\boldsymbol{w}}_1\right)\right| \\
& + \left[c_1 + \frac{(1-\zeta)}{T}\sigma^2 \|\widehat{\boldsymbol{w}}_1\|^2\right] \left|\frac{1}{n}\sum_{i=1}^n \boldsymbol{v}^T \boldsymbol{\gamma}_i\right| \\
& + \left[c_1 |R_{\boldsymbol{v},\boldsymbol{q}_\star}| + c_1 |R_{\boldsymbol{v},\boldsymbol{w}_\star}|\right] \left|\frac{1}{n}\sum_{i=1}^n y_i \left(\boldsymbol{\gamma}_i^T \widehat{\boldsymbol{w}}_1\right)\right| \\
& + \left[c_1\right] \left|\frac{1}{n}\sum_{i=1}^n y_i \left(\boldsymbol{v}^T \boldsymbol{\gamma}_i\right)\right| \\
& + \zeta |R_{\boldsymbol{v},\boldsymbol{q}_\star}| \left|\frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{\gamma}_i^T \widehat{\boldsymbol{w}}_1\right)^2 - \frac{(1-\zeta)}{T}\widehat{\boldsymbol{w}}_1^T \boldsymbol{\Sigma} \widehat{\boldsymbol{w}}_1\right| \\
& + \zeta |R_{\boldsymbol{v},\boldsymbol{w}_\star}| \left|\frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{\gamma}_i^T \widehat{\boldsymbol{w}}_1\right)^2 y_i\right| \\
& + \left[1 + c_1\right] \left|\frac{1}{n}\sum_{i=1}^n y_i \left(\widehat{\boldsymbol{w}}_1^T \boldsymbol{\gamma}_i\right)\left(\boldsymbol{v}^T \boldsymbol{\gamma}_i\right)\right| \\
& + \left[1 + c_1\right] \left|\frac{1}{n}\sum_{i=1}^n y_i \boldsymbol{v}^T \hat{\boldsymbol{\Sigma}}_i \widehat{\boldsymbol{w}}_1\right| \\
& + c_1 \left|\frac{1}{n}\sum_{i=1}^n \boldsymbol{v}^T \hat{\boldsymbol{\Sigma}}_i \widehat{\boldsymbol{w}}_1 - (1-\zeta)\boldsymbol{v}^T \boldsymbol{\Sigma} \widehat{\boldsymbol{w}}_1\right| \\
& + \left[c_1\right] \left|\frac{1}{n}\sum_{i=1}^n \left(\widehat{\boldsymbol{w}}_1^T \boldsymbol{\gamma}_i\right)\left(\boldsymbol{v}^T \boldsymbol{\gamma}_i\right) - \frac{1-\zeta}{T}\boldsymbol{v}^T \boldsymbol{\Sigma} \widehat{\boldsymbol{w}}_1\right| \\
& + \left|\frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{v}^T \boldsymbol{\gamma}_i\right)\left(\left(\widehat{\boldsymbol{w}}_1^T \boldsymbol{\gamma}_i\right)^2 - \frac{(1-\zeta)}{T}\widehat{\boldsymbol{w}}_1^T \boldsymbol{\Sigma} \widehat{\boldsymbol{w}}_1\right)\right| \\
& + \left|\frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{\gamma}_i^T \widehat{\boldsymbol{w}}_1\right)\left(\boldsymbol{v}^T \hat{\boldsymbol{\Sigma}}_i \widehat{\boldsymbol{w}}_1 - (1-\zeta)\widehat{\boldsymbol{w}}_1^T \boldsymbol{\Sigma} \boldsymbol{v}\right)\right| . \qquad (42)
\end{aligned}
$$

Now, we use successively Lemma 5 to bound the random terms. Also note that $|R_{\boldsymbol{v},\boldsymbol{q}_\star}| \leq Q\|\boldsymbol{v}\|$, $|R_{\boldsymbol{v},\boldsymbol{w}_\star}| \leq W\|\boldsymbol{v}\|$ and $\|\widehat{\boldsymbol{w}}_1\| \leq (3/2)\|\boldsymbol{w}_1\| = (3/2)\eta\zeta W$. For any $u > 0$, with probability at least $1 - c' e^{-cu^{2/3}}$ we have

$$
\begin{aligned}
&\left|\boldsymbol{v}^T \widetilde{\boldsymbol{G}}_{\boldsymbol{q}}(0, \widehat{\boldsymbol{w}}_1)\right| \leq u \cdot \frac{C}{\sqrt{n}} \|\boldsymbol{v}\| \, c_1 \, (W+Q) \\
&+ u \cdot \frac{C\sigma\sqrt{1-\zeta}}{\sqrt{nT}} \|\boldsymbol{v}\| \left(c_1 \left(2W + 2Q + \eta\zeta(1-\zeta)\sigma^2 W\right) \eta\zeta W + \left(2c_1 + \sigma^2(1-\zeta)\eta^2\zeta^2 W^2/T\right)\right) \\
&+ u \cdot \frac{C\sigma^2(1-\zeta)}{T\sqrt{n}} \|\boldsymbol{v}\| \left(\eta^2\zeta^3 W^2 Q + \eta^2\zeta^3 W^3 + (1+c_1)\eta\zeta W\right) \\
&+ u \cdot \frac{C\sigma^2\sqrt{1-\zeta}}{\sqrt{nT}} \|\boldsymbol{v}\| \left((1+2c_1)\eta\zeta W\right) \\
&+ u \, \frac{C\sigma^2(1-\zeta)}{T\sqrt{n}} \|\boldsymbol{v}\| \left(c_1 \eta\zeta W\right) \\
&+ u \, \frac{C\sigma^3(1-\zeta)^{3/2}\log(n)}{T^{3/2}\sqrt{n}} \|\boldsymbol{v}\| \left(\eta^2\zeta^2 W^2\right) \\
&+ u \, \frac{C\sigma^3(1-\zeta)\log(n)}{T\sqrt{n}} \|\boldsymbol{v}\| \, \eta^2\zeta^2 W^2 \, . \qquad (43)
\end{aligned}
$$

Now, again using small step size η as per (35), this can be further simplified to the following (here the value of constant c_1 might be different from (43))

$$
\begin{aligned}
\left|\boldsymbol{v}^T \widetilde{\boldsymbol{G}}_{\boldsymbol{q}}(0, \widehat{\boldsymbol{w}}_1)\right| &\leq u \cdot \frac{C}{\sqrt{n}} \|\boldsymbol{v}\| \, c_1 \, (W+Q) + u \cdot \frac{C\sigma\sqrt{1-\zeta}}{\sqrt{nT}} \|\boldsymbol{v}\| \, c_1 \, (1+W+Q) + u \cdot \frac{C\sigma^2(1-\zeta)}{T\sqrt{n}} \|\boldsymbol{v}\| \, c_1 \\
&\quad + u \cdot \frac{C\sigma^2\sqrt{1-\zeta}}{\sqrt{nT}} \|\boldsymbol{v}\| \, c_1 + u \, \frac{C\sigma^3(1-\zeta)^{3/2}\log(n)}{T^{3/2}\sqrt{n}} \|\boldsymbol{v}\| \, c_1 + u \, \frac{C\sigma^3(1-\zeta)\log(n)}{T\sqrt{n}} \|\boldsymbol{v}\| \, c_1 \\
&\leq u \cdot \|\boldsymbol{v}\| \cdot \frac{C}{\sqrt{n}} \cdot c_1 \left(W \vee Q \vee \frac{\sigma \vee \sigma^2}{\sqrt{T}} \vee \frac{\sigma^2}{T} \vee \frac{\sigma^3\log(n)}{T^{3/2}} \vee \frac{\sigma^3\log(n)}{T}\right) \\
&\leq u \cdot \|\boldsymbol{v}\| \cdot \frac{C\log(n)}{\sqrt{n}} \cdot c_1 \left(W \vee Q \vee \frac{\sigma \vee \sigma^2}{\sqrt{T}} \vee \frac{\sigma^3}{T}\right) \qquad (44)
\end{aligned}
$$

Now, we can compute the deviation of the relevance scores. For signal tokens we have for both $y \in \{\pm 1\}$, and all $u > 0$ with probability at least $1 - c'e^{-cu^{2/3}}$, there exist constant $C > 0$ such that

$$
\left|(\boldsymbol{q}_\star + y\boldsymbol{w}_\star)^T \widetilde{\boldsymbol{G}}_{\boldsymbol{q}}(0, \widehat{\boldsymbol{w}}_1)\right| \leq u\, C(W^2 \vee Q^2) \left(1 \vee \frac{\sigma \vee \sigma^2}{\sqrt{T}} \vee \frac{\sigma^3}{T}\right) \frac{\log(n)}{\sqrt{n}} \qquad (45)
$$

For a noisy token $\boldsymbol{z}_t, t \in \mathcal{R}^c$ note that $\boldsymbol{z}_t^T \widetilde{\boldsymbol{G}}_{\boldsymbol{q}}(0, \widehat{\boldsymbol{w}}_1) \sim \mathcal{SN}(\sigma \|\widetilde{\boldsymbol{G}}_{\boldsymbol{q}}(0, \widehat{\boldsymbol{w}}_1)\|)$. We can bound $\|\widetilde{\boldsymbol{G}}_{\boldsymbol{q}}(0, \widehat{\boldsymbol{w}}_1)\|$ applying (44) for all standard basis vectors $\boldsymbol{v} = \boldsymbol{e}_j, j \in [n]$ and union bounding. This gives for all $u > 0$ with probability at least $1 - c'de^{-cu^{2/3}}$,

$$
\|\widetilde{\boldsymbol{G}}_{\boldsymbol{q}}(0, \widehat{\boldsymbol{w}}_1)\| \leq u\, C \left(W \vee Q \vee \frac{\sigma \vee \sigma^2}{\sqrt{T}} \vee \frac{\sigma^3}{T}\right) \frac{\log(n)}{\sqrt{n}} \, .
$$

□

With lemmas 8 and 9 at hand, we are now ready to put things together stating our final bounds for relevance scores. The finding is presented as a stand-alone lemma below.

Lemma 10 (Put things together). *Consider the finite-sample gradient step $\widehat{\boldsymbol{q}}_1 = \widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0, \widehat{\boldsymbol{w}}_1)$, where recall that $\widehat{\boldsymbol{w}}_1 = \eta \widehat{\boldsymbol{G}}_{\boldsymbol{w}}(0, 0)$. Fix any $u_0, u_1, u_2, u_3 > 0$ and any small constant $\widetilde{c}_1 > 0$. Suppose step-size η of first gradient step satisfies (35) for sufficiently small constant $c_\eta > 0$ (dependent on $\widetilde{c}_1$) and further assume*

$$
\sqrt{n} \geq u_0 \cdot C_0 \frac{Q}{W} \qquad \textit{and} \qquad \sqrt{\frac{n\zeta T}{d}} \geq u_0 \cdot C_0 \frac{\sigma}{W} \sqrt{\zeta^{-1} - 1}. \qquad (46)
$$

for some large enough constant $C_0 > 0$. Finally, assume for simplicity that

$$\frac{\sigma \vee \sigma^2}{\sqrt{T}} \vee \frac{\sigma^3}{T} \leq W \vee Q \tag{47}$$

Consider a fresh dataset $(\boldsymbol{X}_i, y_i)_{i\in[n]}$ and consider the signal and noise relevance scores $\hat{r}_{i,t} := r_{i,0} := \left(\boldsymbol{q}_\star + y_i \boldsymbol{w}_\star\right)^T \widehat{\boldsymbol{q}}_1, t \in \mathcal{R}$ and $\hat{r}_{i,t} := \boldsymbol{z}_{i,t}^T \widehat{\boldsymbol{q}}_1, t \in \mathcal{R}^c$, respectively. Then, there exist positive absolute constants $C, c_i, c_i', i = 0, 1, 2, 3$ such that the following statements hold

- *With probability at least $1 - c_0' e^{-c_0 u_0^2} - c_1' e^{c_1 u_1^{2/3}}$, the signal relevance scores satisfy*

$$\min_{i\in[n]} r_{i,0} \geq \eta\zeta(\zeta - \zeta^2) W^2 Q \left(\left(15 - 8\rho^2 - 6|\rho|\right) Q - \left(210|\rho| + 33\right) W\right)/64 - u_1 C\widetilde{c}_1 \left(W^2 \vee Q^2\right) \frac{\log(n)}{\sqrt{n}} =: B(u_1) \tag{48}$$

- *With probability at least $1 - c_0' e^{-c_0 u_0^2} - c_2' n(1-\zeta) T e^{-c_2 u_2^2} - c_3' dn(1-\zeta) T e^{c_3 u_3^{2/3}}$, the noise relevance scores satisfy*

$$\max_{i\in[n], t\in\mathcal{R}^c} r_{i,t} \leq u_2 \eta\zeta \left(\zeta - \zeta^2\right) W^2 Q \sigma C + u_3 \sigma C \widetilde{c}_1 \left(W \vee Q\right) \frac{\log(n)}{\sqrt{n}} =: M(u_2, u_3) \tag{49}$$

Proof. The lemma follows by collecting (36), (39), (37), (40), and applying union bound for the noise terms over $i \in [n], t \in \mathcal{R}^c$. □

F.4 Finishing the finite sample analysis: Third gradient step

With the characterization of the relevant scores from Lemma 10 in hand we are now ready to turn our attention to the third gradient step and finish the finite sample analysis. To this aim we first formally state our finite sample result before turning our attention to its proof. We note that Theorem 6 stated in the main paper follows immediately from this theorem as discussed in Appendix F.4.1.

Theorem 7. *Consider Algorithm 3 with step size η obeying*

$$\eta \leq \frac{c_\eta}{\max\left(Q + W, cu_4\sigma\sqrt{\alpha\log(ndT)}\right)\max\left(Q + W, cu_4\sigma\sqrt{d}\right)\zeta}$$

for small enough constant c_η and with γ sufficiently large. Furthermore, assume the sample size n obeys

$$\sqrt{n} \geq u_0 \cdot C_0 \frac{Q}{W} \qquad \text{and} \qquad \sqrt{\frac{n\zeta T}{d}} \geq u_0 \cdot C_0 \frac{\sigma}{W}\sqrt{\zeta^{-1} - 1}.$$

for large enough constant $C_0 > 0$ and some $u_0 > 0$. For simplicity we also assume

$$\frac{\sigma \vee \sigma^2}{\sqrt{T}} \vee \frac{\sigma^3}{T} \leq W \vee Q\,.$$

Recall (48), (49) and suppose that $B(u_1) \geq 2M(u_2, u_3)$. Also define

$$\delta := \frac{2}{n} e^{-\frac{n}{8}} + (1-\zeta) T e^{-u_4 d} + c_0' e^{-c_0 u_0^2} + c_1' e^{c_1 u_1^{2/3}} + c_2'(1-\zeta) T e^{-c_2 u_2^2} + c_3' d(1-\zeta) T e^{c_3 u_3^{2/3}}$$

Then, we can conclude that

$$\mathrm{ERR}(f_{\boldsymbol{\theta}}') \leq \delta := \frac{2}{n} e^{-\frac{n}{8}} + (1-\zeta) T e^{-u_4 d} + c_0' e^{-c_0 u_0^2} + c_1' e^{c_1 u_1^{2/3}} + c_2'(1-\zeta) T e^{-c_2 u_2^2} + c_3' d(1-\zeta) T e^{c_3 u_3^{2/3}}$$

holds with probability at least

$$\begin{aligned} &1 - 2e^{cu_5^2} - (1-\zeta) T e^{-cu_4 d} - \frac{nT(d+2)}{(ndT)^\alpha} - 2(d+1)e^{-u_6^2} \\ &\quad - c_0' e^{-c_0 u_0^2} - c_1' e^{c_1 u_1^{2/3}} - c_2' n(1-\zeta) T e^{-c_2 u_2^2} - c_3' dn(1-\zeta) T e^{c_3 u_3^{2/3}} \\ &\quad - 2n\left(\frac{2}{n} e^{-\frac{n}{8}} + (1-\zeta) T e^{-cu_4 d} + c_0' e^{-c_0 u_0^2} + c_1' e^{c_1 u_1^{2/3}} + c_2'(1-\zeta) T e^{-c_2 u_2^2} + c_3' d(1-\zeta) T e^{c_3 u_3^{2/3}}\right). \end{aligned}$$

Above we remark that u_7 is free variables. Because it can be chosen arbitrarily large by picking a sufficiently large γ and consequently, its associated failure events can be made arbitrarily small.

F.4.1 Proving Theorem 6 as a Corollary

Setting variables in light of Theorem 7: The log factors will be hidden under $\tilde{\mathcal{O}}()$ notation. $\sigma \propto 1$ and set γ to sufficiently large in terms of all other variables. Set $Q \gtrsim 1 + W$ and $\zeta \leq 0.9$ so that $1 - \zeta \gtrsim 1$. Set $\eta \propto 1/Q^2\zeta$ and $u_4 \propto Q/\sqrt{d}$ and $\alpha \propto \tilde{\mathcal{O}}(d)$ (sufficiently large constant). α results in a failure rate of $1 - e^{-\tilde{\mathcal{O}}(d)}$ however this will only appear in the probability of failure of the training (and not the classification error). In contrast, u_4 results in an error rate of

$$e^{-\tilde{\mathcal{O}}(Q\sqrt{d})} := Te^{-cu_4 d}.$$

For variables u_1, u_3, (48) and (49) yield that

$$\eta\zeta^2 W^2 Q^2 \gtrsim (u_1 Q^2 + u_3 Q)\log(n)/\sqrt{n}$$

Now, set $u_1 \propto u_3 \propto \eta\sqrt{n}\zeta^2 W^2/\log(n)$ and plug in η. This yields an error rate of

$$e^{-\tilde{\mathcal{O}}(n^{1/3}\zeta^{2/3}(W/Q)^{4/3})} := e^{-\tilde{\mathcal{O}}(\eta^{2/3}n^{1/3}\zeta^{4/3}W^{4/3})}.$$

Secondly, we satisfy the u_2 term in (49) via $u_2 \propto Q$ which yields another $e^{-\tilde{\mathcal{O}}(Q^2)}$ rate. These satisfy all conditions of Theorem 7 with the exception of condition for u_0. To proceed, u_0 is chosen to be

$$u_0^2 \propto n[(W/Q)^2 \wedge \zeta^2 W^2 T/d],$$

resulting in an error rate of

$$e^{-n[(W/Q)^2 \wedge \zeta^2 W^2 T/d]}.$$

Combining these, we find

$$\log\delta \leq -\tilde{\mathcal{O}}(Q(\sqrt{d}\wedge Q) \wedge n^{1/3}\zeta^{2/3}(W/Q)^{4/3} \wedge n(W/Q)^2 \wedge n\zeta^2 W^2 T/d) \tag{50}$$

$$= -\tilde{\mathcal{O}}(Q(\sqrt{d}\wedge Q) \wedge n^{1/3}\zeta^{2/3}(W/Q)^{4/3} \wedge n\zeta^2 W^2 T/d). \tag{51}$$

To proceed, note that, the failure probability over the dataset is $cn \times \delta + e^{-\tilde{\mathcal{O}}(d)}$, thus the advertised failure probability (under $\tilde{\mathcal{O}}()$ which subsumes $\times n$) applies to the training process.

In contrast, recall that the linear model achieves a log-error rate of $\log \mathcal{Q}(\sqrt{\zeta^2 T/(1-\zeta)})\|\boldsymbol{w}_\star\| \propto -\zeta^2 TW^2$.

F.4.2 Proof of Theorem 7

Note that under the assumption of the theorem, Lemma 10 holds. For convenience we shall use the notation in that lemma. We also define

$$M(u_2,u_3) := u_2\,\eta\zeta\left(\zeta-\zeta^2\right)W^2 Q\,\sigma\,C + u_3\,\sigma\,C\widetilde{c}_1\,(W\vee Q)\,\frac{\log(n)}{\sqrt{n}}$$

$$B(u_1) := \eta\zeta(\zeta-\zeta^2)W^2Q\left(\left(15-8\rho^2-6|\rho|\right)Q-(210|\rho|+33)W\right)/64 - u_1\,C\widetilde{c}_1\,(W^2\vee Q^2)\frac{\log(n)}{\sqrt{n}}$$

As per theorem assumption, note that $B \geq 2M$.

We define the events

$$\mathcal{E}_i := \left\{\max_{t\in\mathcal{R}^c} r_{i,t} \leq M\right\} \cap \left\{\min_{t\in\mathbb{R}} r_{i,t} \geq \frac{B}{2}\right\}$$

Also note that using the theorem assumption that $B \geq 2M$ we have

$$\begin{aligned}
\frac{\zeta T e^{\gamma r_{i,0}}}{\zeta T e^{\gamma r_{i,0}} + \sum_{t\in\mathcal{R}^c} e^{\gamma r_{i,t}}} &= \frac{1}{1 + \frac{1}{\zeta T}\sum_{t\in\mathcal{R}^c} e^{\gamma(r_{i,t}-r_{i,0})}} \\
&\geq \frac{1}{1 + \frac{1}{\zeta T}\sum_{t\in\mathcal{R}^c} e^{\gamma(\max_{t\in\mathcal{R}^c} r_{i,t}-B)}} \\
&\geq \frac{1}{1 + \frac{1}{\zeta T}\sum_{t\in\mathcal{R}^c} e^{-\frac{\gamma}{2}B}} \\
&= \frac{1}{1 + \frac{1-\zeta}{\zeta} e^{-\frac{\gamma}{2}B}}
\end{aligned}$$

Thus

$$\begin{aligned}\mathbb{1}_{\mathcal{R}_i^c}^T\phi(\boldsymbol{X}_i\boldsymbol{q}) =& 1-\frac{\zeta T e^{\gamma r_{i,0}}}{\zeta T e^{\gamma r_{i,0}}+\sum_{t\in\mathcal{R}^c} e^{\gamma r_{i,t}}}\\ \leq& \frac{\frac{1-\zeta}{\zeta}e^{-\frac{\gamma}{2}B}}{1+\frac{1-\zeta}{\zeta}e^{-\frac{\gamma}{2}B}}\\ =& \frac{1-\zeta}{\zeta e^{\frac{\gamma}{2}B}+(1-\zeta)} := \epsilon_\gamma\end{aligned}$$

The latter also implies that

$$\left|\mathbb{1}_{\mathcal{R}_i}^T\phi(\boldsymbol{X}_i\boldsymbol{q})-1\right| = \mathbb{1}_{\mathcal{R}_i^c}^T\phi(\boldsymbol{X}_i\boldsymbol{q}) \leq \epsilon_\gamma \tag{52}$$

Note that

$$\widehat{\boldsymbol{G}}_{\boldsymbol{w}}(\boldsymbol{q},\boldsymbol{w}) := -\nabla_{\boldsymbol{w}}\widehat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i\in[n]}(y_i - f_{\boldsymbol{\theta}}(\boldsymbol{X}_i))\boldsymbol{X}_i^\top\phi((\boldsymbol{X}_i\boldsymbol{q})).$$

Thus

$$\begin{aligned}\boldsymbol{w}_2 := \widehat{\boldsymbol{G}}_{\boldsymbol{w}}(\boldsymbol{q},\hat{\boldsymbol{w}}_1) =& \frac{1}{n}\sum_{i\in[n]} y_i\boldsymbol{X}_i^\top\phi(\boldsymbol{X}_i\boldsymbol{q}) - \frac{1}{n}\sum_{i\in[n]}\left(\widehat{\boldsymbol{w}}_1^T\boldsymbol{X}_i^\top\phi(\boldsymbol{X}_i\boldsymbol{q})\right)\boldsymbol{X}_i^\top\phi(\boldsymbol{X}_i\boldsymbol{q})\\ =& \frac{1}{n}\sum_{i\in[n]} y_i\left(\boldsymbol{q}_\star + y_i\boldsymbol{w}_\star\right)\mathbb{1}_{\mathcal{R}_i}^T\phi(\boldsymbol{X}_i\boldsymbol{q}) + \frac{1}{n}\sum_{i\in[n]} y_i\boldsymbol{Z}_i^\top\phi(\boldsymbol{X}_i\boldsymbol{q})\\ &-\frac{1}{n}\sum_{i\in[n]}\left(\widehat{\boldsymbol{w}}_1^T\boldsymbol{X}_i^\top\phi(\boldsymbol{X}_i\boldsymbol{q})\right)\boldsymbol{X}_i^\top\phi(\boldsymbol{X}_i\boldsymbol{q})\\ =& \frac{1}{n}\sum_{i\in[n]}(y_i\boldsymbol{q}_\star+\boldsymbol{w}_\star) + \frac{1}{n}\sum_{i\in[n]}(y_i\boldsymbol{q}_\star+\boldsymbol{w}_\star)\left(\mathbb{1}_{\mathcal{R}_i}^T\phi(\boldsymbol{X}_i\boldsymbol{q})-1\right) + \frac{1}{n}\sum_{i\in[n]} y_i\boldsymbol{Z}_i^\top\phi(\boldsymbol{X}_i\boldsymbol{q})\\ &-\frac{1}{n}\sum_{i\in[n]}\left(\widehat{\boldsymbol{w}}_1^T\boldsymbol{X}_i^\top\phi(\boldsymbol{X}_i\boldsymbol{q})\right)\boldsymbol{X}_i^\top\phi(\boldsymbol{X}_i\boldsymbol{q})\\ =& \left(\frac{1}{n}\sum_{i\in[n]} y_i\right)\boldsymbol{q}_\star + \boldsymbol{w}_\star + \frac{1}{n}\sum_{i\in[n]}(y_i\boldsymbol{q}_\star+\boldsymbol{w}_\star)\left(\mathbb{1}_{\mathcal{R}_i}^T\phi(\boldsymbol{X}_i\boldsymbol{q})-1\right) + \frac{1}{n}\sum_{i\in[n]} y_i\boldsymbol{Z}_i^\top\phi(\boldsymbol{X}_i\boldsymbol{q})\\ &-\frac{1}{n}\sum_{i\in[n]}\left(\widehat{\boldsymbol{w}}_1^T\boldsymbol{X}_i^\top\phi(\boldsymbol{X}_i\boldsymbol{q})\right)\boldsymbol{X}_i^\top\phi(\boldsymbol{X}_i\boldsymbol{q})\end{aligned}$$

Let us focus on bounding the penultimate term in a unit norm direct $\boldsymbol{v}$. That is

$$\frac{1}{n}\sum_{i\in[n]} y_i\boldsymbol{v}^T\boldsymbol{Z}_i^\top\phi(\boldsymbol{X}_i\boldsymbol{q}).$$

To this aim we define the events

$$\mathcal{E}_{\boldsymbol{v},i} := \left\{\|\boldsymbol{Z}_i\boldsymbol{v}\|_{\ell_\infty} \leq \sigma\sqrt{c\alpha\log\left(ndT\right)}\right\},$$

and focus on the truncated sum

$$\frac{1}{n}\sum_{i\in[n]} y_i\boldsymbol{v}^T\boldsymbol{Z}_i^\top\phi(\boldsymbol{X}_i\boldsymbol{q})\mathbb{1}_{\mathcal{E}_{\boldsymbol{v},i}}\mathbb{1}_{\mathcal{E}_i}.$$

To continue note that $y_i\boldsymbol{v}^T\boldsymbol{Z}_i^\top\phi(\boldsymbol{X}_i\boldsymbol{q})\mathbb{1}_{\mathcal{E}_{\boldsymbol{v},i}}\mathbb{1}_{\mathcal{E}_i}$ is a sub-Gaussian random variable as it is bounded due to the fact that

$$\left|y_i\boldsymbol{v}^T\boldsymbol{Z}_i^\top\phi(\boldsymbol{X}_i\boldsymbol{q})\mathbb{1}_{\mathcal{E}_{\boldsymbol{v},i}}\mathbb{1}_{\mathcal{E}_i}\right| \leq \|\boldsymbol{Z}_i\boldsymbol{v}\|_{\ell_\infty}\mathbb{1}_{\mathcal{R}_i^c}^T\phi(\boldsymbol{X}_i\boldsymbol{q})\mathbb{1}_{\mathcal{E}_{\boldsymbol{v},i}}\mathbb{1}_{\mathcal{E}_i} \leq \sigma\sqrt{c\alpha\log(ndT)}\epsilon_\gamma$$

Thus by concentration of sum of i.i.d. Sub-Gaussian random variables we have

$$\left|\frac{1}{n}\sum_{i\in[n]} y_i\boldsymbol{v}^T\boldsymbol{Z}_i^\top\phi(\boldsymbol{X}_i\boldsymbol{q})\mathbb{1}_{\mathcal{E}_{\boldsymbol{v},i}}\mathbb{1}_{\mathcal{E}_i} - \mathbb{E}\left[y_i\boldsymbol{v}^T\boldsymbol{Z}_i^\top\phi(\boldsymbol{X}_i\boldsymbol{q})\mathbb{1}_{\mathcal{E}_{\boldsymbol{v},i}}\mathbb{1}_{\mathcal{E}_i}\right]\right| \leq cu_6\sigma\frac{\sqrt{c\alpha\log(ndT)}\epsilon_\gamma}{\sqrt{n}}$$

holds with probability at least $1 - 2e^{-u_6^2}$. Furthermore, note that by Jensen's inequalities

$$\left|\mathbb{E}\Big[y_i \boldsymbol{v}^T \boldsymbol{Z}_i^\top \phi(\boldsymbol{X}_i \boldsymbol{q}) \mathbb{1}_{\mathcal{E}_{\boldsymbol{v},i}} \mathbb{1}_{\mathcal{E}_i}\Big]\right| \le \mathbb{E}\Big[\left|y_i \boldsymbol{v}^T \boldsymbol{Z}_i^\top \phi(\boldsymbol{X}_i \boldsymbol{q}) \mathbb{1}_{\mathcal{E}_{\boldsymbol{v},i}} \mathbb{1}_{\mathcal{E}_i}\right|\Big] \le \mathbb{E}\big[\|Z_i \boldsymbol{v}\|_{\ell_\infty} \mathbb{1}_{\mathcal{R}_i^c}^T \phi(\boldsymbol{X}_i \boldsymbol{q}) \mathbb{1}_{\mathcal{E}_{\boldsymbol{v},i}} \mathbb{1}_{\mathcal{E}_i}\big] \le \sigma\sqrt{c\alpha \log(ndT)}\epsilon_\gamma$$

Now use the fact that

$$\frac{1}{n}\sum_{i\in[n]} y_i \boldsymbol{v}^T \boldsymbol{Z}_i^\top \phi(\boldsymbol{X}_i \boldsymbol{q}) = \frac{1}{n}\sum_{i\in[n]} y_i \boldsymbol{v}^T \boldsymbol{Z}_i^\top \phi(\boldsymbol{X}_i \boldsymbol{q}) \mathbb{1}_{\mathcal{E}_{\boldsymbol{v},i}} \mathbb{1}_{\mathcal{E}_i}$$

holds with probability at least $1 - \frac{nT}{(ndT)^\alpha} - \sum_{i=1}^n \mathbb{P}\{\mathcal{E}_i^c\}$. Combining this with the latter two inequalities (via the triangular inequality) we conclude that

$$\left|\frac{1}{n}\sum_{i\in[n]} y_i \boldsymbol{v}^T \boldsymbol{Z}_i^\top \phi(\boldsymbol{X}_i \boldsymbol{q})\right| \le c\sigma\epsilon_\gamma\sqrt{c\alpha \log(ndT)}\left(\frac{Cu_6}{\sqrt{n}} + 1\right) \tag{53}$$

holds with probability at least $1 - \frac{nT}{(ndT)^\alpha} - \sum_{i=1}^n \mathbb{P}\{\mathcal{E}_i^c\} - 2e^{-u_6^2}$. Applying the above to standard basis vectors $\boldsymbol{e}_j \in \mathbb{R}^d, j \in [d]$ we can conclude that

$$\|\frac{1}{n}\sum_{i\in[n]} y_i \boldsymbol{v}^T \boldsymbol{Z}_i^\top \phi(\boldsymbol{X}_i \boldsymbol{q})\| \le c\sigma\epsilon_\gamma\sqrt{c\alpha \log(ndT)}\left(\frac{Cu_6}{\sqrt{n}} + 1\right)$$

holds with probability at least $1 - \frac{ndT}{(ndT)^\alpha} - 2de^{-u_6^2} - \sum_{i=1}^n \mathbb{P}\{\mathcal{E}_i^c\}$.

To control the last term note that we have

$$\begin{aligned}
\|\frac{1}{n}\sum_{i\in[n]} \left(\widehat{\boldsymbol{w}}_1^T \boldsymbol{X}_i^\top \phi(\boldsymbol{X}_i \boldsymbol{q})\right) \boldsymbol{X}_i^\top \phi(\boldsymbol{X}_i \boldsymbol{q})\| &\le \frac{1}{n}\sum_{i\in[n]} \|\boldsymbol{X}_i \hat{\boldsymbol{w}}_1\|_{\ell_\infty} \left(\max_t \|\boldsymbol{e}_t^T \boldsymbol{X}_i\|\right) \\
&\le \max\left(\|\boldsymbol{q}_\star\| + \|\boldsymbol{w}_\star\|, cu_4\sigma\sqrt{\alpha \log(ndT)}\right) \max\left(\|\boldsymbol{q}_\star\| + \|\boldsymbol{w}_\star\|, cu_4\sigma\sqrt{d}\right) \|\hat{\boldsymbol{w}}_1\| \\
&\le \max\left(\|\boldsymbol{q}_\star\| + \|\boldsymbol{w}_\star\|, cu_4\sigma\sqrt{\alpha \log(ndT)}\right) \max\left(\|\boldsymbol{q}_\star\| + \|\boldsymbol{w}_\star\|, cu_4\sigma\sqrt{d}\right) \frac{3}{2}\eta\zeta W \\
&= \max\left(Q + W, cu_4\sigma\sqrt{\alpha \log(ndT)}\right) \max\left(Q + W, cu_4\sigma\sqrt{d}\right) \frac{3}{2}\eta\zeta W
\end{aligned}$$

holds with probability at least $1 - \frac{nT}{(ndT)^\alpha} - (1-\zeta)Te^{-cu_4 d}$ where in the penultimate line we used (34). Assuming

$$\eta \le \frac{1}{3\max\left(Q + W, cu_4\sigma\sqrt{\alpha \log(ndT)}\right) \max\left(Q + W, cu_4\sigma\sqrt{d}\right) \zeta}$$

the latter is less than $\frac{1}{2}W$.

Using the above together with triangular inequality on $\boldsymbol{w}_2$

$$\begin{aligned}
\|\boldsymbol{w}_2\| &\le \left|\frac{1}{n}\sum_{i\in[n]} y_i\right| \|\boldsymbol{q}_\star\| + \|\boldsymbol{w}_\star\| + \left(\|\boldsymbol{q}_\star\| + \|\boldsymbol{w}_\star\|\right)\epsilon_\gamma + c\sigma\epsilon_\gamma\sqrt{c\alpha \log(ndT)}\left(\frac{Cu_6}{\sqrt{n}} + 1\right) + \frac{W}{2} \\
&\le \frac{u_5}{\sqrt{n}}\|\boldsymbol{q}_\star\| + \|\boldsymbol{w}_\star\| + \left(\|\boldsymbol{q}_\star\| + \|\boldsymbol{w}_\star\|\right)\epsilon_\gamma + c\sigma\epsilon_\gamma\sqrt{c\alpha \log(ndT)}\left(\frac{Cu_6}{\sqrt{n}} + 1\right) + \frac{W}{2} \\
&= \frac{u_5}{\sqrt{n}}Q + W + \left(Q + W\right)\epsilon_\gamma + c\sigma\epsilon_\gamma\sqrt{c\alpha \log(ndT)}\left(\frac{Cu_6}{\sqrt{n}} + 1\right) + \frac{W}{2}
\end{aligned}$$

holds with probability at least $1 - (1-\zeta)Te^{-cu_4 d} - 2e^{cu_5^2} - \frac{nT(d+1)}{(ndT)^\alpha} - 2de^{-u_6^2} - \sum_{i=1}^n \mathbb{P}\{\mathcal{E}_i^c\}$.

Furthermore, using (53) we also have

$$\left|\frac{1}{n}\sum_{i\in[n]} y_i \boldsymbol{w}_\star^T \boldsymbol{Z}_i^\top \phi(\boldsymbol{X}_i \boldsymbol{q})\right| \le c\sigma\epsilon_\gamma\sqrt{c\alpha \log(ndT)}\left(\frac{Cu_6}{\sqrt{n}} + 1\right)\|\boldsymbol{w}_\star\|$$

with an added failure probability of $\frac{nT}{(ndT)^\alpha} + 2e^{-u_6^2}$. Therefore,

$$\begin{aligned}\boldsymbol{w}_2^T\boldsymbol{w}_\star \geq& \|\boldsymbol{w}_\star\|^2 - \frac{u_5}{\sqrt{n}}\left|\boldsymbol{q}_\star^T\boldsymbol{w}_\star\right| - \left(\left|\boldsymbol{q}_\star^T\boldsymbol{w}_\star\right| + \|\boldsymbol{w}_\star\|^2\right)\epsilon_\gamma - c\sigma\epsilon_\gamma\sqrt{c\alpha\log(ndT)}\left(\frac{Cu_6}{\sqrt{n}}+1\right)\|\boldsymbol{w}_\star\| - \frac{W^2}{2}\\ =&(\frac{1}{2}-\epsilon_\gamma)W^2 - \left(\frac{u_5}{\sqrt{n}}+\epsilon_\gamma\right)QW|\rho| - c\sigma\epsilon_\gamma\sqrt{c\alpha\log(ndT)}\left(\frac{Cu_6}{\sqrt{n}}+1\right)W\end{aligned}$$

Combining the upper bound on $\|\boldsymbol{w}_2\|$ and lower bound on $\boldsymbol{w}_2^T\boldsymbol{w}_\star$ we can thus conclude that

$$\frac{\boldsymbol{w}_2^T\boldsymbol{w}_\star}{\|\boldsymbol{w}_2\|} = \frac{(\frac{1}{2}-\epsilon_\gamma)W^2 - \left(\frac{u_5}{\sqrt{n}}+\epsilon_\gamma\right)QW|\rho| - c\sigma\epsilon_\gamma\sqrt{c\alpha\log(ndT)}\left(\frac{Cu_6}{\sqrt{n}}+1\right)W}{\frac{u_5}{\sqrt{n}}Q + \frac{3}{2}W + (Q+W)\epsilon_\gamma + c\sigma\epsilon_\gamma\sqrt{c\alpha\log(ndT)}\left(\frac{Cu_6}{\sqrt{n}}+1\right)}$$

with probability at least

$$\begin{aligned}&1 - 2e^{cu_5^2} - \frac{nT(d+2)}{(ndT)^\alpha} - 2(d+1)e^{-u_6^2} - \sum_{i=1}^n \mathbb{P}\{\mathcal{E}_i^c\}\\ &= 1 - 2e^{cu_5^2} - (1-\zeta)Te^{-cu_4 d} - \frac{nT(d+2)}{(ndT)^\alpha} - 2(d+1)e^{-u_6^2}\\ &\quad - c_0'e^{-c_0u_0^2} - c_1'e^{c_1u_1^{2/3}} - c_0'e^{-c_0u_0^2} - c_2'n(1-\zeta)Te^{-c_2u_2^2} - c_3'dn(1-\zeta)Te^{c_3u_3^{2/3}}.\end{aligned}$$

Finally note that

$$\begin{aligned}\|\boldsymbol{X}^T\phi(\boldsymbol{X}\boldsymbol{q}) - (\boldsymbol{q}_\star + y\boldsymbol{w}_\star)\| =& \|(\boldsymbol{q}_\star + y\boldsymbol{w}_\star)\left(\mathbb{1}_{\mathcal{R}}^T\phi(\boldsymbol{X}\boldsymbol{q}) - 1\right) + \boldsymbol{Z}^T\phi(\boldsymbol{X}\boldsymbol{q})\|\\ \leq& (Q + W + \max_t\|\boldsymbol{z}_t\|)\epsilon_\gamma\\ \leq& (Q + W + cu_4\sigma\sqrt{d})\epsilon_\gamma := \epsilon\end{aligned}$$

which holds with probability at least $1 - (1-\zeta)Te^{-cu_4d} - c_0'e^{-c_0u_0^2} - c_1'e^{c_1u_1^{2/3}} - c_0'e^{-c_0u_0^2} - c_2'(1-\zeta)Te^{-c_2u_2^2} - c_3'd(1-\zeta)Te^{c_3u_3^{2/3}}$.

Note that by picking

$$\delta := \frac{2}{n}e^{-\frac{n}{8}} + (1-\zeta)Te^{-cu_4d} - c_0'e^{-c_0u_0^2} + c_1'e^{c_1u_1^{2/3}} + c_0'e^{-c_0u_0^2} + c_2'(1-\zeta)Te^{-c_2u_2^2} + c_3'd(1-\zeta)Te^{c_3u_3^{2/3}}$$

we also automatically have $n \geq 8\log\left(\frac{2}{\delta n}\right)$. Thus by picking γ sufficiently large the assumptions of Lemma 11 holds with δ and ϵ as defined above. Therefore, we can conclude that

$$\text{ERR}(f_{\boldsymbol{\theta}}') \leq \delta := \frac{2}{n}e^{-\frac{n}{8}} + (1-\zeta)Te^{-cu_4d} + c_0'e^{-c_0u_0^2} + c_1'e^{c_1u_1^{2/3}} + c_0'e^{-c_0u_0^2} + c_2'(1-\zeta)Te^{-c_2u_2^2} + c_3'd(1-\zeta)Te^{c_3u_3^{2/3}}$$

holds with probability at least

$$\begin{aligned}&1 - 2e^{cu_5^2} - (1-\zeta)Te^{-cu_4d} - \frac{nT(d+2)}{(ndT)^\alpha} - 2(d+1)e^{-u_6^2}\\ &\quad - c_0'e^{-c_0u_0^2} - c_1'e^{c_1u_1^{2/3}} - c_0'e^{-c_0u_0^2} - c_2'n(1-\zeta)Te^{-c_2u_2^2} - c_3'dn(1-\zeta)Te^{c_3u_3^{2/3}}\\ &\quad - 2n\left(\frac{2}{n}e^{-\frac{n}{8}} + (1-\zeta)Te^{-cd} + c_0'e^{-c_0u_0^2} + c_1'e^{c_1u_1^{2/3}} + c_0'e^{-c_0u_0^2} + c_2'(1-\zeta)Te^{-c_2u_2^2} + c_3'd(1-\zeta)Te^{c_3u_3^{2/3}}\right)\end{aligned}$$

F.4.3 De-biasing step

Lemma 11 (Debiasing predictions)**.** *Suppose $\boldsymbol{q}_1$ is such that a test example $(y, \boldsymbol{X})$ obeys*

$$\mathbb{P}(\|\boldsymbol{X}^\top\phi(\boldsymbol{X}\boldsymbol{q}_1) - (\boldsymbol{q}_\star + y\boldsymbol{w}_\star)\| \leq \epsilon) \geq 1 - \delta.$$

Given a fresh dataset $\mathcal{S} = (y_i, \boldsymbol{X}_i)_{i=1}^n$, set $b = \frac{1}{n}\sum_{i=1}^n f_{\boldsymbol{\theta}}(\boldsymbol{X}_i) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{w}_2^\top\boldsymbol{v}_i$ where $\boldsymbol{v}_i := \boldsymbol{X}_i^\top\phi(\boldsymbol{X}_i\boldsymbol{q}_1)$. Set the debiased classifier $f_{\boldsymbol{\theta}}'(\boldsymbol{X}) = f_{\boldsymbol{\theta}}(\boldsymbol{X}) - b$. Set $\Delta = \boldsymbol{w}_2^\top\boldsymbol{w}_\star/\|\boldsymbol{w}_2\|$. Suppose $n \geq 8\log\left(\frac{2}{\delta n}\right)$ and $\epsilon > 0$ is such that $\epsilon < \Delta/4$. Then, with probability $1 - 2\delta n$ over $\mathcal{S}$, the test error of $f_{\boldsymbol{\theta}}'$ obeys

$$\text{ERR}(f_{\boldsymbol{\theta}}') \leq \delta.$$

Proof. First, let us prove the following intermediate statement: With probability $1-2\delta n$ over $\mathcal{S}$, for a new test sample $(y, \boldsymbol{X})$, with probability $1-\delta$,

$$|yf'_{\boldsymbol{\theta}}(\boldsymbol{X}) - \boldsymbol{w}_2^\top \boldsymbol{w}_\star| \leq \sqrt{\frac{2\log(2/\delta n)}{n}} \cdot \boldsymbol{w}_2^\top \boldsymbol{w}_\star + 2\epsilon\|\boldsymbol{w}_2\|. \tag{54}$$

Based on the choice of C, observe that, with probability $1-n\delta$ over the dataset $(y_i, \boldsymbol{X}_i)_{i=1}^n$, for each $\boldsymbol{v}_i$,

$$|\boldsymbol{w}_2^\top \boldsymbol{v}_i - \boldsymbol{w}_2^\top(\boldsymbol{q}_\star + y_i \boldsymbol{w}_\star)| \leq \epsilon\|\boldsymbol{w}_2\|.$$

Set $\bar{b} = \boldsymbol{w}_2^\top \boldsymbol{q}_\star$. Set $\bar{y} = |\frac{1}{n}\sum_{i=1}^n y_i|$. With probability $1-\delta n$, $\bar{y} \leq \sqrt{\frac{2\log(2/\delta n)}{n}}$. Combining, with overall probability at least $1-2\delta n$, the classifier bias obeys

$$|b - \bar{b}| \leq \left|\boldsymbol{w}_2^\top \boldsymbol{w}_\star\right| \sqrt{\frac{2\log(2/\delta n)}{n}} + \epsilon\|\boldsymbol{w}_2\|.$$

To finalize, for a new sample $(y, \boldsymbol{X})$, with probability $1-\delta$, we have that $|\boldsymbol{w}_2^\top \boldsymbol{v} - \boldsymbol{w}_2^\top(\boldsymbol{q}_\star + y\boldsymbol{w}_\star)| \leq \epsilon\|\boldsymbol{w}_2\|$ where $\boldsymbol{v} = \boldsymbol{X}^\top \phi(\boldsymbol{X}\boldsymbol{q}_1)$. Thus, the prediction $f'(\boldsymbol{X}) = f(\boldsymbol{X}) - b$ obeys

$$|yf'_{\boldsymbol{\theta}}(\boldsymbol{X}) - y(f_{\boldsymbol{\theta}}(\boldsymbol{X}) - \bar{b})| \leq |b - \bar{b}| \leq \left|\boldsymbol{w}_2^\top \boldsymbol{w}_\star\right| \sqrt{\frac{2\log(2/\delta n)}{n}} + \epsilon\|\boldsymbol{w}_2\|. \tag{55}$$

To conclude with (54), note that

$$|y(f_{\boldsymbol{\theta}}(\boldsymbol{X}) - \bar{b}) - \boldsymbol{w}_2^\top \boldsymbol{w}_\star| \leq \epsilon\|\boldsymbol{w}_2\|,$$

and apply triangle inequality with (55).

To prove the statement of the theorem, note that, when $n \geq 8\log(2/\delta n)$ and $\boldsymbol{w}_2^\top \boldsymbol{w}_\star > 4\epsilon\|\boldsymbol{w}_2\|$, a test sample (with $\geq 1-\delta$ probability) obeys

$$yf'_{\boldsymbol{\theta}}(\boldsymbol{X}) \geq \boldsymbol{w}_2^\top \boldsymbol{w}_\star - \sqrt{\frac{2\log(2/\delta n)}{n}} \boldsymbol{w}_2^\top \boldsymbol{w}_\star - 2\epsilon\|\boldsymbol{w}_2\| \geq 0.5\boldsymbol{w}_2^\top \boldsymbol{w}_\star - 2\epsilon\|\boldsymbol{w}_2\| > 0. \tag{56}$$

Thus, classifier makes the correct decision with the same probability. □

G Proofs for Population-Gradient Analysis in Section E.2

This section includes the missing proofs of all the results in Section E.2 regarding population analysis of Algorithm 3.

G.1 Proof of Lemma 2

We repeat here the lemma for convenience also stated for general (not necessarily isotropic) noise covariance $\boldsymbol{\Sigma}$.

Lemma 12. *The second population gradient step $\boldsymbol{q}_1 = \gamma \boldsymbol{G}_{\boldsymbol{w}}(\boldsymbol{w}_1, 0)$ satisfies the following for $\alpha := \eta\zeta$*

$$\begin{aligned}\boldsymbol{G}_{\boldsymbol{q}}(0, \alpha\boldsymbol{w}_\star) = &\left((\zeta-\zeta^2)\left(\alpha W^2 + \alpha^2 \boldsymbol{w}_\star^T \boldsymbol{\Sigma}\boldsymbol{w}_\star/T\right) - \alpha^2(\zeta^2-\zeta^3)\left(W^4 + \left(\boldsymbol{w}_\star^T \boldsymbol{q}_\star\right)^2\right)\right)\boldsymbol{q}_\star \\ &+ \left(\left((\zeta-\zeta^2) - 2(\zeta^2-\zeta^3)\alpha W^2\right)\alpha\left(\boldsymbol{w}_\star^T \boldsymbol{q}_\star\right)\right)\boldsymbol{w}_\star \\ &- \left((1+2/T)(\zeta-\zeta^2)\alpha\left(\boldsymbol{w}_\star^T \boldsymbol{q}_\star\right)\right)\alpha\boldsymbol{\Sigma}\boldsymbol{w}_\star \end{aligned} \tag{57}$$

Proof. The lemma follows immediately from Eqn. (18) of Lemma 5 by recognizing that for $\boldsymbol{w} = \alpha\boldsymbol{w}_\star$ it holds $R_{\boldsymbol{q}_\star} = \alpha \boldsymbol{q}_\star^T \boldsymbol{w}_\star$ and $R_{\boldsymbol{w}_\star} = \alpha W^2$. □

G.2 COROLLARY 1

Corollary 1. *Suppose small enough step-size η obeying*

$$\eta\left(\zeta^2\left(W^2+Q^2\right)-\sigma^2/T\right)\le 1/2. \tag{58a}$$
$$\eta\zeta\left(2\zeta W^2+(1+2/T)\sigma^2\right)\le 5/4. \tag{58b}$$
$$\eta\zeta\left(\sigma^2/T\right)\le 1/2. \tag{58c}$$

Then, for $C_1\in[1/2,3/2]$ and $C_2\in[-1/4,1]$, we have that

$$\boldsymbol{q}_1=\gamma\eta\zeta(\zeta-\zeta^2)W\left(C_1W\boldsymbol{q}_\star+C_2\rho Q\boldsymbol{w}_\star\right).$$

In particular, $\boldsymbol{q}_\star^\top\boldsymbol{q}_1=\gamma\eta\zeta(\zeta-\zeta^2)W^2Q^2\left(C_1+C_2\rho^2\right)$ and $\boldsymbol{w}_\star^\top\boldsymbol{q}_1=\gamma\eta\zeta(\zeta-\zeta^2)W^3Q\rho\left(C_1+C_2\right)$.

Proof. Set $\alpha=\eta\zeta$ and

$$3/2\ge C_1:=\left(1+\alpha\sigma^2/T\right)-\alpha\zeta\left(W^2+\rho^2Q^2\right)\ge 1/2. \tag{59a}$$
$$1\ge C_2:=1-2\alpha\zeta W^2-(1+2/T)\alpha\sigma^2\ge -1/4. \tag{59b}$$

The gradient formula follows directly from (4). For the lower/upper bounds on C_1,C_2 use (58a), (58b) and (58c). □

Remark 1 (Condition on correlation)**.** *To classify correctly the signal tokens, we need*

$$y\boldsymbol{w}_\star^T(\boldsymbol{q}_\star+y\boldsymbol{w}_\star)>0\quad\Longleftrightarrow\quad y\rho Q+W>0\Longleftrightarrow\quad|\rho|<W/Q \tag{60}$$

Note that if (60) holds, then

$$C_1\ge 1+\alpha\sigma^2/T-2\alpha\zeta W^2=C_2+(1+3/T)\alpha\sigma^2. \tag{61}$$

G.3 PROOF OF THEOREM 5

From Corollary 1, we have $\boldsymbol{q}_1^\gamma=\gamma\eta\zeta(\zeta-\zeta^2)W\left(C_1W\boldsymbol{q}_\star+C_2\rho Q\boldsymbol{w}_\star\right)$ for $3/2\ge C_1\ge 1/2$ and $1\ge C_2\ge -1/4$. Set $\boldsymbol{a}=\phi(\boldsymbol{X}\boldsymbol{q}_1^\gamma)$. Our goal is to compare the relevance scores $r_t=\boldsymbol{z}_t^T\boldsymbol{q}_1^\gamma, t\in\mathcal{R}^c$ of noisy tokens to the relevance scores $r_t=(\boldsymbol{q}_\star+y\boldsymbol{w}_\star)^T\boldsymbol{q}_1^\gamma, t\in\mathcal{R}$ of signal tokens:

$$\begin{aligned}t\in\mathcal{R}:\quad r_t=(\boldsymbol{q}_\star+y\boldsymbol{w}_\star)^T\boldsymbol{q}_1^\gamma&=\gamma\eta\zeta(\zeta-\zeta^2)W\left((C_1+C_2\rho^2)WQ^2+y\rho(C_1+C_2)W^2Q\right)\\&=\gamma\eta\zeta(\zeta-\zeta^2)W^2Q\left((C_1+C_2\rho^2)Q+y\rho(C_1+C_2)W\right)=:\gamma B_y,\end{aligned} \tag{62}$$

where the notation B_y emphasizes the dependency on the label y (unless $\rho=0$), and,

$$\begin{aligned}t\in\mathcal{R}^c:\quad r_t:=\boldsymbol{z}_t^T\boldsymbol{q}_1^\gamma&=\gamma\eta\zeta(\zeta-\zeta^2)W\left(C_1W\boldsymbol{q}_\star^T\boldsymbol{z}_t+\rho C_2Q\boldsymbol{w}_\star^T\boldsymbol{z}_t\right)\\&=\gamma\eta\zeta(\zeta-\zeta^2)W^2Q\left(C_1\overline{\boldsymbol{q}}_\star^T\boldsymbol{z}_t+\rho C_2\overline{\boldsymbol{w}}_\star^T\boldsymbol{z}_t\right)\\&\overset{D}{=}\gamma\eta\zeta(\zeta-\zeta^2)W^2Q\sigma\sqrt{2}\sqrt{C_1^2+\rho^2C_2(C_2+2C_1)}\cdot G_t,\quad G_t\overset{\text{IID}}{\sim}\mathcal{SN}(1)\\&=:\gamma\Gamma,\end{aligned} \tag{63}$$

where we have denoted for convenience $G_t:=\frac{C_1\overline{\boldsymbol{q}}_\star^T\boldsymbol{z}_t+\rho C_2\overline{\boldsymbol{w}}_\star^T\boldsymbol{z}_t}{\sigma\sqrt{2}\sqrt{C_1^2+\rho^2C_2(C_2+2C_1)}}\sim\mathcal{SN}(1)$.

With these, the attention coefficients $a_t:=\boldsymbol{a}[t]$ are

$$t\in\mathcal{R}:\quad a_t=e^{\gamma B_y}/S_y=:a_{0,y} \tag{64a}$$
$$t\in\mathcal{R}^c:\quad a_t=e^{\gamma\Gamma G_t}/S_y. \tag{64b}$$

where we denote $S_y:=\zeta Te^{\gamma B_y}+\sum_{t\in\mathcal{R}^c}e^{\gamma\Gamma G_t}$ and $a_{0,y}:=e^{\gamma B_y}/S_y$ for convenience.

Thus,

$$\boldsymbol{w}_2 := \boldsymbol{G}_{\boldsymbol{w}}(0, \boldsymbol{q}_1^\gamma) = \mathbb{E}\left[y\boldsymbol{X}^T\phi(\boldsymbol{X}\boldsymbol{q}_1^\gamma)\right] = \mathbb{E}\left[y\boldsymbol{X}^T\boldsymbol{a}\right]$$
$$= \mathbb{E}\left[\left(\frac{\zeta T e^{\gamma B_y}}{\zeta T e^{\gamma B_y} + \sum_{t\in\mathcal{R}^c} e^{\gamma\Gamma G_t}}\right)(\boldsymbol{w}_\star + y\boldsymbol{q}_\star)\right] + \sum_{t\in\mathcal{R}^c}\mathbb{E}\left[\frac{e^{\gamma\Gamma G_t}}{\zeta T e^{\gamma B_y} + \sum_{t\in\mathcal{R}^c} e^{\gamma\Gamma G_t}} y\boldsymbol{z}_t\right] \tag{65}$$

Denote $M := \sqrt{2\log(2(1-\zeta)T^\alpha)}$, define the good event

$$\mathcal{E} := \left\{\max_{t\in\mathcal{R}^c}|G_t| \le M\right\} \tag{66}$$

and note that[2]

$$\Pr(\mathcal{E}) \ge 1 - 1/T^{\alpha-1} =: 1-\delta.$$

In this event, we show that (7) implies

$$\Gamma\cdot\max_{t\in\mathcal{R}^c} G_t \le \frac{1}{2}\min_{y\in\{\pm1\}} B_y. \tag{67}$$

To see this, observe that the desired (67) holds under $\mathcal{E}$ provided

$$M\sigma\sqrt{2}\sqrt{C_1^2 + \rho^2 C_2(C_2+2C_1)} \le \frac{1}{2}\left((C_1 + C_2\rho^2)Q - |\rho|(C_1+C_2)W\right) \tag{68}$$

Now use $C_1 > 0$, as well as, $C_2 \le C_1$ from (61) for the lower bound) to note that $\sqrt{C_1^2 + \rho^2 C_2(C_2+2C_1)} \le C_1\sqrt{1+3\rho^2}$. Hence, it suffices that

$$C\sigma\sqrt{\log((1-\zeta)T)} \le \frac{1 + (C_2/C_1)\rho^2}{2\sqrt{2}\sqrt{1+3\rho^2}}Q - \frac{|\rho|(1+C_2/C_1)}{2\sqrt{2}\sqrt{1+3\rho^2}}W$$

Using $C_2/C_1 \in [-1/2, 1]$ (for the lower bound recall $C_2 \ge -1/4, C_1 \ge 1/2$) we arrive at (7).

Now, note the following implications of (67) (i.e. holding conditioned on the good event $\mathcal{E}$). First, for all $y \in \{\pm1\}$,

$$T(\zeta e^{B_y\gamma} + (1-\zeta)e^{B_y\gamma/2}) \ge S_y \ge T\zeta e^{B_y\gamma}.$$

Using this, we can show similar to the proof for the finite-sample case in Section F.4.2 that for sufficiently large γ, the attention weights attend up to a small error to the context-relevant tokens (see Eqn. (52)). To avoid repeating those approximation arguments, we assume onwards for simplicity that $\gamma \to \infty$.

Then, conditioned on $\mathcal{E}$, as $\gamma\to\infty$, the sequence of random variables $\Psi_{\mathcal{R}}^\gamma = \frac{e^{B_y\gamma}}{S_y}$ converges pointwise (over the sample space of random y and $\boldsymbol{z}_t, t\in\mathcal{R}^c$) to $1/(\zeta T)$, i.e. $\Psi_{\mathcal{R}}^\gamma = \frac{e^{B_y\gamma}}{S_y} \to 1/(\zeta T)$. Similarly, for all $t\in\mathcal{R}^c$ the sequences $\Psi_{\mathcal{R}^c,t}^\gamma = \frac{e^{\gamma\Gamma G_t}}{S_y} \to 0$ because from (67): $e^{\gamma\Gamma G_t} < e^{\gamma B_y/2}$ and $\frac{e^{\gamma B_y/2}}{S_y}\to 0$. Thus, conditioned on $\mathcal{E}$ the following pointwise convergence is true:

$$a_t^\gamma := \boldsymbol{a}_t = \begin{cases} = \frac{e^{B_y\gamma}}{S_y} \to \frac{1}{T\zeta} & \text{if} \quad t\in\mathcal{R}, \\ \le \frac{e^{B_y\gamma/2}}{S_y} \to 0 & \text{if} \quad t\in\mathcal{R}^c. \end{cases}, \tag{69}$$

where the added superscript γ denotes the (random variables of) attention weights are parameterized by γ.

We now use (69) towards computing the limit of $yf_{\boldsymbol{\theta}}$ for $\boldsymbol{\theta} = (\boldsymbol{w}_2, \boldsymbol{q}_1^\gamma)$. Recall from (65) that

$$\boldsymbol{w}_2^\gamma = \sum_{t\in\mathcal{R}}\mathbb{E}\left[a_t^\gamma(\boldsymbol{w}_\star + y\boldsymbol{q}_\star)\right] + \sum_{t\in\mathcal{R}^c}\mathbb{E}\left[a_t^\gamma y\boldsymbol{z}_t\right]. \tag{70}$$

Thus, it suffices to study the limits of the following sequences of random variables:

$$A_{\boldsymbol{w}_\star}^\gamma := \boldsymbol{w}_2^T\boldsymbol{w}_\star, \quad A_{\boldsymbol{q}_\star}^\gamma := \boldsymbol{w}_2^T\boldsymbol{q}_\star, \quad \text{and} \quad A_{\tilde{\boldsymbol{z}}_t}^\gamma := \boldsymbol{w}_2^T\tilde{\boldsymbol{z}}_t$$

[2]Recall $G_t \sim \mathcal{SN}(1)$. Thus, by union bound $\mathbb{P}(\max_t|G_t| > M) \le 2(1-\zeta)Te^{-M^2/2}$.

since the model output at $\boldsymbol{\theta} = (\boldsymbol{w}_2^\gamma, \boldsymbol{q}_1^\gamma)$ is

$$y f_{\boldsymbol{\theta}^\gamma}(\boldsymbol{X}) = \sum_{t\in\mathcal{R}} \tilde{a}_t^\gamma \left(A^\gamma_{\boldsymbol{w}_\star} + y A^\gamma_{\boldsymbol{q}_\star}\right) + \sum_{t\in\mathcal{R}^c} \tilde{a}_t^\gamma y A^\gamma_{\tilde{\boldsymbol{z}}_t} \tag{71}$$

and we use $\tilde{a}_t^\gamma$ to denote the attention weights evaluated at test data $(y, \boldsymbol{X})$. This distingushing them from a_t^γ in (70), but note (69) still holds for $\tilde{a}_t^\gamma$.

We start by controlling $A^\gamma_{\boldsymbol{w}_\star}$. We have

$$A^\gamma_{\boldsymbol{w}_\star} = \mathbb{E}\left[\sum_{t\in\mathcal{R}} a_t^\gamma (W^2 + y\rho QW)\right] + \mathbb{E}\left[\sum_{t\in\mathcal{R}^c} a_t^\gamma y \boldsymbol{w}_\star^T \boldsymbol{z}_t\right]. \tag{72}$$

Denote $\Phi^\gamma_{\mathcal{R}}$ and $\Phi^\gamma_{\mathcal{R}^c}$ the two random variables inside the expectations in the above display. Recall that $|a_t^\gamma| \le 1$ and $|y| \le 1$. This implies $\Phi^\gamma_{\mathcal{R}}$ is absolutely bounded, i.e. $|\Phi^\gamma_{\mathcal{R}}| \le \zeta T\left(W^2 + |\rho| QW\right)$. Thus, applying dominated convergence theorem (DCT) and using (69) together with $\mathbb{E}[y] = 0$:

$$\lim_{\gamma\to\infty} \mathbb{E}\left[\Phi^\gamma_{\mathcal{R}} \,|\mathcal{E}\right] = \mathbb{E}\left[\lim_{\gamma\to\infty} \Phi^\gamma_{\mathcal{R}} \,|\mathcal{E}\right] = \mathbb{E}\left[W^2 + y\rho QW\right] = W^2.$$

Moreover, since $\sum_{t\in\mathcal{R}} a_t^\gamma \le 1$ and $W^2 + y\rho QW \ge 0$,

$$\mathbb{E}\left[\Phi^\gamma_{\mathcal{R}} | \mathcal{E}^c\right] = \mathbb{E}\left[\mathbb{E}_y\left[\left(W^2 + y\rho QW\right)\sum_{t\in\mathcal{R}} a_t^\gamma\right] | \mathcal{E}^c\right] \le \mathbb{E}\left[\mathbb{E}_y\left[\left(W^2 + y\rho QW\right)\right] | \mathcal{E}^c\right] \le W^2\,.$$

Combining the above two displays, we find

$$\mathbb{E}\left[\Phi^\gamma_{\mathcal{R}}\right] \le \mathbb{E}\left[\Phi^\gamma_t \,|\mathcal{E}\right] + \delta \mathbb{E}\left[\Phi^\gamma_t\right] \le (1+\delta)\, W^2\,. \tag{73}$$

Consider now $\Phi^\gamma_{\mathcal{R}^c} = \sum_{t\in\mathcal{R}^c} a_t^\gamma y \boldsymbol{w}_\star^T \boldsymbol{z}_t$. Note that

$$\Phi^\gamma_{\mathcal{R}^c} = \sum_{t\in\mathcal{R}^c} a_t^\gamma y \boldsymbol{z}_t^T \boldsymbol{w}_\star \le W \max_{t\in\mathcal{R}^c} \|\boldsymbol{z}_t\| \sum_{t\in\mathcal{R}^c} a_t^\gamma \le W \max_{t\in\mathcal{R}^c} \|\boldsymbol{z}_t\|,$$

where we used that $a_t^\gamma \ge 0$, $|y| \le 1$ and $\sum_{t\in\mathcal{R}^c} a_t^\gamma \le 1$. Thus, $\Phi^\gamma_{\mathcal{R}^c}$ is absolutely bounded since $\mathbb{E}[\max_{t\in\mathcal{R}^c} \|\boldsymbol{z}_t\|] < \infty$; thus, by DCT and (69)

$$\lim_{\gamma\to\infty} \mathbb{E}\left[\Phi^\gamma_{\mathcal{R}^c} \,|\mathcal{E}\right] = \mathbb{E}\left[\lim_{\gamma\to\infty} \Phi^\gamma_{\mathcal{R}^c} \,|\mathcal{E}\right] = 0\,.$$

Next, we bound $\mathbb{E}\left[\Phi^\gamma_{\mathcal{R}^c} | \mathcal{E}^c\right]\mathbb{P}(\mathcal{E}^c)$as follows: $\mathbb{E}\left[\Phi^\gamma_{\mathcal{R}^c} | \mathcal{E}^c\right]\mathbb{P}(\mathcal{E}^c) \le W\,\mathbb{E}\left[\max_{t\in\mathcal{R}^c} \|\boldsymbol{z}_t\| | \mathcal{E}^c\right]\mathbb{P}(\mathcal{E}^c)$ and use Lemma 13 to conclude that

$$\mathbb{E}\left[\Phi^\gamma_{\mathcal{R}^c} | \mathcal{E}^c\right]\mathbb{P}(\mathcal{E}^c) \le 12 W\delta\sigma\sqrt{d}\sqrt{\log\left(2(1-\zeta)T/\delta\right)}\,.$$

The above two displays combined yield:

$$\lim_{\gamma\to\infty} \mathbb{E}\left[\Phi^\gamma_{\mathcal{R}^c}\right] \le 12 W\delta\sigma\sqrt{d}\sqrt{\log\left(2(1-\zeta)T/\delta\right)}\,.$$

Putting things together, we have shown that

$$\lim_{\gamma\to\infty} A^\gamma_{\boldsymbol{w}_\star} \le (1+\delta)\, W^2 + CW\delta\sigma\sqrt{d}\sqrt{\log\left(2(1-\zeta)T/\delta\right)}. \tag{74}$$

In exact same way we can show that

$$\lim_{\gamma\to\infty} A^\gamma_{\boldsymbol{q}_\star} \le \rho WQ + \delta\left(|\rho| WQ + Q^2\right) + CQ\delta\sigma\sqrt{d}\sqrt{\log\left(2(1-\zeta)T/\delta\right)}\,, \tag{75}$$

and for all $t \in \mathcal{R}^c$

$$\lim_{\gamma\to\infty} A^\gamma_{\tilde{\boldsymbol{z}}_t} \le \boldsymbol{w}_\star^T \tilde{\boldsymbol{z}}_t + \delta\left(W + |\rho| Q\right)\|\tilde{\boldsymbol{z}}_t\| + C\|\tilde{\boldsymbol{z}}_t\|\delta\sigma\sqrt{d}\sqrt{\log\left(2(1-\zeta)T/\delta\right)}\,. \tag{76}$$

To continue, we can now use the above displays together with the limits of attention coefficients in (69) to compute from (71) that on the event $\tilde{\mathcal{E}}$ (over the randomness of $\tilde{z}_t$):

$$\begin{aligned}\lim_{\gamma\to\infty} yyf_{\boldsymbol{\theta}^\gamma}(\boldsymbol{X}) =& W(W+y\rho Q)\\ &+\left(W^2+yQ^2+y|\rho|QW\right)/T^{\alpha-1}\\ &+C(W+yQ)\sigma\sqrt{d}\sqrt{\log\left(2(1-\zeta)T^{2-\alpha}\right)}\Big/T^{\alpha-1}\end{aligned}$$

Note that (60) together with the following

$$W(W-\rho Q)\geq\frac{C}{T^{\alpha-1}}\left(\left(-W^2+Q^2+|\rho|QW\right)+(-W+Q)\sigma\sqrt{d}\sqrt{\log\left(2(1-\zeta)T^{2-\alpha}\right)}\right) \tag{77}$$

guarantee the expression on the RHS above is nonnegative. That is, under event $\mathcal{E}$, $\lim_{\gamma\to\infty}\hat{y}^\gamma>0$ Hence,

$$\text{ERR}_\infty\leq\Pr\left(\lim_{\gamma\to\infty} yf_{\boldsymbol{\theta}^\gamma}(\boldsymbol{X})<0\,\middle|\,\mathcal{E}\right)+\Pr(\mathcal{E}^c)\leq 1/T^{\alpha-1}.$$

G.3.1 Auxiliary lemma

Lemma 13 (Subgaussian euclidean-norm tail control)**.** *Let Let $\boldsymbol{z}_i\in\mathbb{R}^d, i\in[N]$ be K-subgaussian random vectors. Then, for any event $\mathcal{E}$ with $\mathbb{P}(\mathcal{E})=\delta$, it holds that*

$$\mathbb{E}\left[\max_{i\in[n]}\|\boldsymbol{z}_i\|\,\middle|\,\mathcal{E}^c\right]\leq 12K\sqrt{d}\sqrt{\log(2N/\delta)}.$$

Proof. Set $Z=\max_{i\in[n]}\|\boldsymbol{z}_i\|$ and define event $\mathcal{B}=\{Z\geq M\}$ for $M:=4K\sqrt{d}\sqrt{\log(2N/\delta)}$. By Fact J.4 for all $t>0$, $\mathbb{P}(Z>t)\leq 2Ne^{-t^2/(16dK^2)}$. Thus, by choice of M, $\mathbb{P}(\mathcal{B})\leq\mathbb{P}(\mathcal{E})=\delta$.

Denote the pdf and cdf complement of Z by f_Z, Q_Z respectively. Observe that, we set $Q_Z(M)\leq\delta$. Using integration by parts we have,

$$\begin{aligned}\mathbb{E}[Z|\mathcal{B}]\mathbb{P}(\mathcal{B}) &= \int_M^\infty zf_Z(z)\mathrm{d}z=-\int_M^\infty zdQ_Z(z)=\int_M^\infty Q_Z(z)\mathrm{d}z-[Q_Z(z)z]_M^\infty\\ &=\int_M^\infty Q_Z(z)\mathrm{d}z+\delta M\\ &=\delta M+\int_M^\infty\mathbb{P}(Z\geq t)\mathrm{d}t\leq\delta M+\int_M^\infty 2Ne^{-t^2/(16dK^2)}\mathrm{d}t\\ &\leq\delta M+2\sqrt{2}K\sqrt{d}(2N)\int_{\sqrt{2\log(2N/\delta)}}^\infty e^{-u^2/2}\mathrm{d}u\\ &=\delta\,4K\sqrt{d}\sqrt{\log(2N/\delta)}+2\sqrt{\pi}K\sqrt{d}\,\delta\leq 2\delta M.\end{aligned}$$

We can conclude the proof by noting:

$$\begin{aligned}\mathbb{E}\left[Z\middle|\mathcal{E}\right]\mathbb{P}(\mathcal{E}) &= \mathbb{E}[Z|\mathcal{E}\cap\mathcal{B}^c]\mathbb{P}(\mathcal{E}\cap\mathcal{B}^c)+\mathbb{E}[Z|\mathcal{E}\cap\mathcal{B}]\mathbb{P}(\mathcal{E}\cap\mathcal{B})\\ &\leq M\delta+\mathbb{E}[Z|\mathcal{B}]\mathbb{P}(\mathcal{B}).\end{aligned}$$

□

H Proofs of the Results on Discrete Datasets

H.1 Proof of Theorem 1 and Observation 1

• **Proof for Prompt Attention:** Let $\bar{\boldsymbol{w}}_\star=\boldsymbol{w}_\star/\|\boldsymbol{w}_\star\|$ and $\bar{\boldsymbol{q}}_\star=\boldsymbol{q}_\star/\|\boldsymbol{q}_\star\|$. $\boldsymbol{q}'_\star$ be the projection of $\boldsymbol{q}_\star$ to the orthogonal complement of $\boldsymbol{w}_\star$ i.e. $\boldsymbol{q}'_\star=\boldsymbol{q}_\star-\bar{\boldsymbol{w}}_\star\bar{\boldsymbol{w}}_\star^\top\boldsymbol{q}_\star$. Similarly, let $\boldsymbol{w}'_\star$ be the projection of $\boldsymbol{w}_\star$ to the orthogonal complement of $\boldsymbol{q}_\star$ i.e. $\boldsymbol{w}'_\star=\boldsymbol{w}_\star-\bar{\boldsymbol{q}}_\star\bar{\boldsymbol{q}}_\star^\top\boldsymbol{w}_\star$. Denote correlation coefficient between two vectors by $\rho(\boldsymbol{a},\boldsymbol{b})=\frac{\boldsymbol{a}^\top\boldsymbol{b}}{\|\boldsymbol{a}\|\|\boldsymbol{b}\|}$.

To proceed, observe that, $\boldsymbol{q}_\star'^\top \boldsymbol{q}_\star = \|\boldsymbol{q}_\star\|^2 - (\bar{\boldsymbol{w}}_\star^\top \boldsymbol{q}_\star)^2 = \|\boldsymbol{q}_\star\|^2(1-\rho(\boldsymbol{q}_\star,\boldsymbol{w}_\star)^2) > 0$. The positivity follows from the fact that $\boldsymbol{q}_\star, \boldsymbol{w}_\star$ are not parallel, thus, the absolute value of their correlation coefficient is strictly bounded away from 1. Similarly $\boldsymbol{w}_\star'^\top \boldsymbol{w}_\star = \|\boldsymbol{w}_\star\|^2(1-\rho(\boldsymbol{q}_\star,\boldsymbol{w}_\star)^2)>0$. To proceed, set $\bar{\rho} := 1-\rho(\boldsymbol{q}_\star,\boldsymbol{w}_\star)^2$ and observe that the classifier $\boldsymbol{\theta} = (\boldsymbol{w}_\star', \Gamma \boldsymbol{q}_\star')$ achieves the attention scores

$$\boldsymbol{a}_i = \phi(\boldsymbol{X}\boldsymbol{q}_\star')_i = \begin{cases} S^{-1} e^{\|\boldsymbol{q}_\star\|^2 \Gamma \bar{\rho}} & \text{if} \quad i \text{ relevant} \\ S^{-1} e^{-\|\boldsymbol{q}_\star\|^2 \Gamma \delta^q \bar{\rho}} & \text{if} \quad i \text{ irrelevant} \end{cases}.$$

where $S = T\zeta e^{\|\boldsymbol{q}_\star\|^2\Gamma\bar{\rho}} + T(1-\zeta)e^{-\|\boldsymbol{q}_\star\|^2\Gamma\delta^q\bar{\rho}}$. Using orthogonality of $\boldsymbol{w}_\star'$ and $\boldsymbol{q}_\star$, the final prediction obeys

$$y f_{\boldsymbol{\theta}}(\boldsymbol{X}) = \|\boldsymbol{w}_\star\|^2 \bar{\rho} S^{-1}\left[\zeta e^{\|\boldsymbol{q}_\star\|^2\Gamma\bar{\rho}} - \delta^w(1-\zeta)e^{-\|\boldsymbol{q}_\star\|^2\Gamma\delta^q\bar{\rho}}\right]$$

The classifier achieves perfect accuracy when $\zeta e^{\|\boldsymbol{q}_\star\|^2\Gamma\bar{\rho}} > |\delta^w|(1-\zeta)e^{-\|\boldsymbol{q}_\star\|^2\Gamma\delta^q\bar{\rho}}$. Since we have $\delta^q \geq 0$ and we have assumed δ^w is a C-bounded variable (i.e. $|\delta^w| \leq C$), thus, the desired inequality can be guaranteed by choosing

$$\Gamma > \frac{1}{\|\boldsymbol{q}_\star\|^2\bar{\rho}} \log\left(\frac{C(1-\zeta)}{\zeta}\right).$$

• **Proof for Observation 1:** To prove this, observe that for any $\delta^q = \Delta^q$, $\delta^w = \Delta^w$ choices, using orthogonality of $\boldsymbol{q}_\star, \boldsymbol{w}_\star$, for any $(y, \boldsymbol{X}) \sim \mathcal{D}$, we have

$$y f^{\text{LIN}}(\boldsymbol{w}_\star') = \|\boldsymbol{w}_\star\|^2 \bar{\rho}(\zeta - (1-\zeta)\delta^w).$$

Thus, as long as $\delta^w \neq \zeta/(1-\zeta)$, $\texttt{sign}(y f^{\text{LIN}}(\boldsymbol{w}_\star'))$ is always 1 or always -1, resulting in perfect accuracy for $\boldsymbol{w}_\star'$ or $-\boldsymbol{w}_\star'$.

• **Proof for Self-attention:** The proof is provided under Theorem 8.

• **Proof for Linear Prompt Attention:** Let $W_1 = \boldsymbol{w}^\top \boldsymbol{w}_\star, W_2 = \boldsymbol{w}^\top \boldsymbol{q}_\star$, $Q_1 = \boldsymbol{q}^\top \boldsymbol{q}_\star, Q_2 = \boldsymbol{q}^\top \boldsymbol{w}_\star$. Since context-irrelevant tokens are of the form $-\delta^q \boldsymbol{q}_\star - y\delta^w \boldsymbol{w}_\star$, the model decision is given by $\frac{1}{T} f(\boldsymbol{X}) = \frac{1}{T}\boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{q} = \zeta \boldsymbol{w}^\top(y\boldsymbol{w}_\star + \boldsymbol{q})(y\boldsymbol{w}_\star + \boldsymbol{q})^\top \boldsymbol{q} + (1-\zeta)\boldsymbol{w}^\top(y\delta^w \boldsymbol{w}_\star + \delta^q \boldsymbol{q})(y\delta^w \boldsymbol{w}_\star + \delta^q \boldsymbol{q})^\top \boldsymbol{q}$ and

$$\frac{1}{T} f(\boldsymbol{X}) = \zeta(yW_1 + W_2)(Q_1 + yQ_2) + (1-\zeta)(y\delta^w W_1 + \delta^q W_2)(y\delta^w Q_2 + \delta^q Q_1) \tag{78}$$

$$= \zeta y(W_1 Q_1 + W_2 Q_2) + \zeta(W_2 Q_1 + W_1 Q_2) + \tag{79}$$

$$(1-\zeta) y\delta^q \delta^w (W_1 Q_1 + W_2 Q_2) + (1-\zeta)(\delta^{q2} W_2 Q_1 + \delta^{w2} W_1 Q_2). \tag{80}$$

$$\frac{y f(\boldsymbol{X})}{T} = (\zeta + (1-\zeta)\delta^q\delta^w)(W_1 Q_1 + W_2 Q_2) + y((\zeta + (1-\zeta)\delta^{q2})W_2 Q_1 + (\zeta + (1-\zeta)\delta^{w2})W_1 Q_2). \tag{81}$$

To proceed, set (δ^q, δ^w) to be $(0,0)$ or $(\Delta, -\Delta)$ equally-likely for $\Delta > \sqrt{\zeta/(1-\zeta)}$. For fixed Δ, for any choice of W_1, W_2, Q_1, Q_2 observe that, with $1/2$ probability the event $E = \{y((\zeta + (1-\zeta)\delta^{q2})W_2 Q_1 + (\zeta + (1-\zeta)\delta^{w2})W_1 Q_2) \leq 0\}$ happens. On this event (which is over the label y), probability that $(\zeta + (1-\zeta)\delta^q\delta^w)(W_1 Q_1 + W_2 Q_2) > 0$ is at most $1/2$ because $\text{sign}(\zeta + (1-\zeta)\delta^q\delta^w)$ is Rademacher variable. Combining, we find that $\mathbb{P}(\frac{y f(\boldsymbol{X})}{T} \leq 0) \geq 25\%$ as advertised whenever $\Delta > \sqrt{\zeta/(1-\zeta)}$.

H.2 Failure Proof for Self-attention

We have the following theorem regarding self-attention.

Theorem 8. *Fix $\Delta > 0$ to be sufficiently large. In (DATA), choose $\delta = (\delta^q, \delta^w)$ to be $(0,0)$ or (Δ, Δ) equally-likely, where $\Delta > 1/(1-\zeta)^2$.*

- *For any choice of $(\boldsymbol{U} = \mathbb{1}\boldsymbol{u}^\top, \boldsymbol{W})$, $f^{\text{SATT}}(\mathbb{1}\boldsymbol{u}^\top, \boldsymbol{W})$ achieves 50% accuracy (i.e. random guess).*
- *For any choice of $(\boldsymbol{U}, \boldsymbol{W})$, there exists a (DATA) distribution with adversarial relevance set choices such that $f^{\text{SATT}}(\boldsymbol{U}, \boldsymbol{W})$ achieves 50% accuracy.*

Here, adversarial relevance set choice means that, the relevance set can be chosen adaptively to the label y, out-of-context term δ, and the self-attention model weights $(\boldsymbol{U}, \boldsymbol{W})$ to cause misclassification.

Proof. Let $\widetilde{\boldsymbol{w}} = \boldsymbol{W}\boldsymbol{w}_\star$ and $\widetilde{\boldsymbol{q}} = \boldsymbol{W}\boldsymbol{q}_\star$. Also let $b_w = \boldsymbol{u}^\top \boldsymbol{w}_\star$ and $b_q = \boldsymbol{u}^\top \boldsymbol{q}_\star$. Since $\boldsymbol{W}$ is allowed to be full-rank and arbitrary, $\widetilde{\boldsymbol{w}}, \widetilde{\boldsymbol{q}}$ are allowed to be arbitrary as well (but fixed given $\boldsymbol{W}$). In our analysis, the critical terms are the attention weights given by the correlation between the relevant/irrelevant keys/queries.

Setting attention queries as the raw tokens (without losing any generality), relevant queries $\boldsymbol{x}_R$ and keys $\boldsymbol{k}_R$ become

$$\boldsymbol{x}_R = y\boldsymbol{w}_\star + \boldsymbol{q}_\star, \ \boldsymbol{k}_R = y\widetilde{\boldsymbol{w}} + \widetilde{\boldsymbol{q}}.$$

Thanks to our choice of $\delta := \delta^w = \delta^q$ to be equally-likely in $\{0, \Delta\}$, observe that irrelevant queries and keys are simply

$$\boldsymbol{x}_I = -\delta\boldsymbol{x}_R, \ \boldsymbol{k}_I = -\delta\boldsymbol{k}_R.$$

This will greatly help the proof because it will mean that attention weights are highly structured. Specifically, set $\rho = \boldsymbol{x}_R^\top \boldsymbol{k}_R$. All weights of the attention similarities belong to the set $(\rho, -\delta\rho, \delta^2\rho)$. Consequently, softmax-attention output $\boldsymbol{A} = \phi(\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^\top)\boldsymbol{X} = \begin{bmatrix} \boldsymbol{a}_1^\top \\ \vdots \\ \boldsymbol{a}_T^\top \end{bmatrix}$ is given by

$$\boldsymbol{a}_i = \begin{cases} \frac{\zeta e^\rho - \delta(1-\zeta)e^{-\delta\rho}}{\zeta e^\rho + (1-\zeta)e^{-\delta\rho}} \cdot \boldsymbol{x}_R & \text{if} \quad i \in \mathcal{R} \text{ (relevant)} \\ \frac{\zeta e^{-\delta\rho} - \delta(1-\zeta)e^{\delta^2\rho}}{\zeta e^{-\delta\rho} + (1-\zeta)e^{\delta^2\rho}} \cdot \boldsymbol{x}_R & \text{if} \quad i \in \mathcal{R}^c \text{ (irrelevant)} \end{cases}. \tag{82}$$

Set $a_+ = \frac{e^\rho}{\zeta e^\rho + (1-\zeta)e^{-\delta\rho}}$, $a_- = \frac{e^{-\delta\rho}}{\zeta e^\rho + (1-\zeta)e^{-\delta\rho}}$, $b_- = \frac{e^{-\delta\rho}}{\zeta e^{-\delta\rho} + (1-\zeta)e^{\delta^2\rho}}$, $b_+ = \frac{e^{\delta^2\rho}}{\zeta e^{-\delta\rho} + (1-\zeta)e^{\delta^2\rho}}$. With this, we also set

$$\Delta_R = \frac{\zeta e^\rho - \delta(1-\zeta)e^{-\delta\rho}}{\zeta e^\rho + (1-\zeta)e^{-\delta\rho}} = \zeta a_+ - \delta(1-\zeta)a_- \tag{83}$$

$$\Delta_I = \frac{\zeta e^{-\delta\rho} - \delta(1-\zeta)e^{\delta^2\rho}}{\zeta e^{-\delta\rho} + (1-\zeta)e^{\delta^2\rho}} = \zeta b_- - \delta(1-\zeta)b_+. \tag{84}$$

Also define $\Delta_i = \Delta_R$ if i is relevant and Δ_I otherwise. With this, we have $\boldsymbol{a}_i = \Delta_i \boldsymbol{x}_r$ based on (82).

The following lemma will be helpful for the downstream analysis. The goal of this lemma is showing that, by choosing $\delta \in \{0, \Delta\}$, we can confuse the model output.

Lemma 14. *Fix a scalar κ. Set $f_\delta = \kappa\Delta_R + (1-\kappa)\Delta_I$. Recalling $\rho = \boldsymbol{x}_R^\top \boldsymbol{k}_R$, the following statements hold:*

- *Set $\delta = 0$. Suppose "$1 \geq \kappa \geq 0$" OR "$\kappa \geq 1, \rho \geq 0$" OR "$\kappa \leq 0, \rho \leq 0$". Then $f_\delta > 0$.*
- *Fix $0 \leq \alpha \leq 1$. Suppose*

$$\delta > \Delta_0 := \frac{1}{1-\zeta} \max\left(\frac{\zeta}{\alpha(1-\zeta)}, \frac{1}{1-\alpha}\right).$$

and "$\kappa \leq \alpha, \rho \geq 0$" OR "$\kappa \geq \alpha, \rho \leq 0$". Then $f_\delta < 0$.

Proof. Plugging in δ, we write

$$f_\delta = \kappa\Delta_R + (1-\kappa)\Delta_I = \kappa\zeta a_+ - \delta\kappa(1-\zeta)a_- + \zeta(1-\kappa)b_- - \delta(1-\zeta)(1-\kappa)b_+ \tag{85}$$

$$= \zeta(\kappa a_+ + (1-\kappa)b_-) - \delta(1-\zeta)(\kappa a_- + (1-\kappa)b_+). \tag{86}$$

• **Suppose $\delta = 0$.** In this case, we obtain the first statement of the lemma as follows

$$f_\delta/\zeta = \frac{\kappa e^\rho}{\zeta e^\rho + 1 - \zeta} + 1 - \kappa > 0 \quad \text{whenever} \quad \begin{cases} 1 \geq \kappa \geq 0 & \text{OR} \\ \kappa \geq 1, \rho \geq 0 & \text{OR} \\ \kappa \leq 0, \rho \leq 0 \end{cases} \tag{87}$$

• **Now suppose $\delta > \Delta_0$. First, assume $\rho \geq 0$ and $\kappa \leq \alpha$.** We use the facts

$$1/(\zeta T) \geq a_+ \geq 1/T, \quad 1/T \geq a_- \geq 0, \quad b_+ \geq 1/T, \quad 1/T \geq b_- \geq 0.$$

Observe that, since $b_+ \geq a_-$ and $\kappa \leq \alpha$

$$\kappa a_- + (1-\kappa) b_+ \geq \begin{cases} b_+ & \text{if} \quad \kappa \leq 0 \\ (1-\alpha) b_+ & \text{if} \quad \kappa \geq 0 \end{cases} \geq (1-\alpha)/T.$$

Additionally, if $\kappa \leq 0$, we have that

$$\kappa a_- + (1-\kappa) b_+ \geq b_+ \geq b_- \geq \kappa a_+ + (1-\kappa) b_-.$$

If $\kappa \leq 0$, we obtain $f_\delta \leq \zeta b_- - \delta(1-\zeta) b_+$. Thus, $T f_\delta < 0$ whenever $\delta > \Delta_0 \geq \zeta/(1-\zeta)$.

If $\kappa \geq 0$, we use $\kappa a_+ + (1-\kappa) b_- \leq 1/(\zeta T)$ to obtain that whenever $\delta > \Delta_0 \geq \frac{1}{(1-\zeta)(1-\alpha)}$

$$T f_\delta \leq 1 - \delta(1-\zeta)(1-\alpha) < 0.$$

Now assume $\rho \leq 0$ and $\kappa \geq \alpha$. We use the facts

$$1/T \geq a_+ \geq 0, \quad a_- \geq 1/T, \quad 1/T \geq b_+ \geq 0, \quad \frac{1}{(1-\zeta)T} \geq b_- \geq 1/T.$$

Observe that, since $b_+ \leq a_-$ and $\kappa \geq \alpha$

$$\kappa a_- + (1-\kappa) b_+ \geq \begin{cases} a_- & \text{if} \quad \kappa \geq 1 \\ \alpha a_- & \text{if} \quad \kappa \leq 1 \end{cases} \geq \alpha/T.$$

Additionally, if $\kappa \geq 1$, we have that

$$\kappa a_- + (1-\kappa) b_+ \geq a_- \geq a_+ \geq \kappa a_+ + (1-\kappa) b_-.$$

If $\kappa \geq 1$, we obtain $f_\delta \leq \zeta a_+ - \delta(1-\zeta) a_-$. Thus, $T f_\delta < 0$ whenever $\delta > \Delta_0 \geq \zeta/(1-\zeta)$.

If $\kappa \leq 1$, we use $\kappa a_+ + (1-\kappa) b_- \leq \frac{1}{(1-\zeta)T}$ to obtain that whenever $\delta > \Delta_0 \geq \frac{\zeta}{(1-\zeta)^2 \alpha}$

$$T f_\delta \leq \frac{\zeta}{1-\zeta} - \delta(1-\zeta)\alpha < 0.$$

□

To proceed, we will conclude with the proof as follows. Set $\nu_i = y \boldsymbol{u}_i^\top \boldsymbol{x}_R$ for $i \in [T]$ where $\boldsymbol{u}_i$ is the ith row of the output layer weights $\boldsymbol{U}$. Here ν_i is obviously y-dependent however we will show that for any choice of y, the model accuracy is at most 50%. Thus we fix y and (mostly) omit it from the notation during the following discussion. Let $\boldsymbol{a}_i$ be the ith token of the attention output. The linear output layer $\boldsymbol{U}$ aggregates $\boldsymbol{u}_i^\top \boldsymbol{a}_i$ to obtain

$$y f(\boldsymbol{U}, \boldsymbol{W}) = \sum_{i=1}^T \boldsymbol{u}_i^\top \boldsymbol{a}_i = \sum_{i=1}^T \nu_i \Delta_i.$$

Aggregating $v_+ = \frac{1}{T} \sum_{i \in \mathcal{R}=\text{relevant}} v_i$ and $v_- = \frac{1}{T} \sum_{i \in \mathcal{R}^c=\text{irrelevant}} v_i$ and recalling from (82) that over relevant/irrelevant sets attention tokens are given by $\Delta_R \boldsymbol{x}_R$ and $\Delta_I \boldsymbol{x}_I$, we find

$$\frac{1}{T} y f(\boldsymbol{U}, \boldsymbol{W}) = \nu_R \Delta_R + \nu_I \Delta_I.$$

Scenario 1: Rows of $\boldsymbol{U}$ are identical and we have $\boldsymbol{U} = \mathbb{1}\boldsymbol{u}^\top$. In this scenario, we simply have $\nu_i = \nu$ and $\nu_R = \zeta\nu$ and $\nu_I = (1-\zeta)\nu$. Thus, we find

$$\frac{1}{T} y f(\boldsymbol{U}, \boldsymbol{W}) = T\nu[\zeta \Delta_R + (1-\zeta)\Delta_I].$$

Set $f_\delta = \zeta \Delta_R + (1-\zeta)\Delta_I$. We claim that $\text{sign}(f_\delta)$ is Rademacher (given arbitrary y choice) which will prove that accuracy is at most 50%. Specifically, let us apply Lemma 14 with $\kappa = \zeta$ and $\alpha = \zeta$. When $\delta = 0$, we have $f_\delta > 0$. When $\delta = \Delta$, since the conditions $\kappa \leq \alpha$ and $\kappa \geq \alpha$ hold, for any choice

of ρ, for $\Delta > \Delta_0 := \frac{1}{(1-\zeta)^2}$ we have that $f_\delta < 0$. Thus, we have that $\mathbb{P}_\delta(f_\delta > 0) = \mathbb{P}_\delta(f_\delta < 0) = 0.5$ as advertised. This follows from the fact that $f_\delta > 0$ for $\delta = 0$ and $f_\delta < 0$ for $\delta = \Delta$ and δ is equally likely over two options.

Scenario 2: Suppose rows of $\boldsymbol{U}$ are not identical. In this case, we will leverage the fact that relevant set $\mathcal{R}$ is allowed to be chosen adversarially with respect to the self-attention weights. We will show that by selecting $\mathcal{R}$ adversarially, on any label y event, accuracy is a coin flip.

First consider the scenario $\nu_{\text{tot}} := \nu_R + \nu_I \le 0$: We will show that model achieves at least 50% error on label y: Let us denote ν_R with $\nu_R^{\mathcal{R}}$ which makes the relevance set dependence explicit. Given $\mathcal{R}$, fixing $\delta = 0$, the model outputs (following (87))

$$\frac{1}{T} y f(\boldsymbol{U}, \boldsymbol{W}) = \frac{\nu_R^{\mathcal{R}} e^{\rho}}{\zeta e^{\rho} + 1 - \zeta} + \nu_I^{\mathcal{R}}.$$

Suppose there is a relevance set $\mathcal{R}_0$ (that depends on y) such that the right hand side is non-positive. Let us select this $\mathcal{R}_0$ as our relevance set. Then, the model makes 50% error on label y thanks to the event $\delta = 0$ (which is exactly what we want). If there is no such $\mathcal{R}_0$, then, for all $\mathcal{R}$, we have

$$\frac{\nu_R^{\mathcal{R}} e^{\rho}}{\zeta e^{\rho} + 1 - \zeta} + \nu_I^{\mathcal{R}} > 0$$

By taking average of all relevance sets ("T choose ζT" many), all v_i's will be equally-weighted and we obtain $\nu_{\text{tot}} = \nu_R + \nu_I > 0$. This contradicts with our initial $\nu_{\text{tot}} \le 0$ assumption, thus, $\mathcal{R}_0$ has to exist.

Now consider the scenario $\nu_{\text{tot}} = \nu_R + \nu_I > 0$: Let $\mathcal{D}$ be the uniform distribution over "T choose ζT" relevant sets $\mathcal{R}$. Clearly $\mathbb{E}_{\mathcal{D}}[\nu_R^{\mathcal{R}}] = \zeta \nu_{\text{tot}} > 0$. Thus, there is a relevance set $\mathcal{R}_+$ such that $\nu_R^{\mathcal{R}_+} \ge \zeta \nu_{\text{tot}}$ and there is a relevance set $\mathcal{R}_-$ such that $\nu_R^{\mathcal{R}_-} \le \zeta \nu_{\text{tot}}$. We will make use of these two sets to finalize the proof.

To proceed, set $\kappa_\pm = \nu^{\mathcal{R}_\pm} / \nu_{\text{tot}}$ and again set $\alpha = \zeta$ and $\Delta_0 = \frac{1}{(1-\zeta)^2}$ in Lemma 14. Here, we are investigating the sign of the prediction

$$\frac{1}{T \nu_{\text{tot}}} y f(\boldsymbol{U}, \boldsymbol{W}) = \frac{\nu_R \Delta_R + \nu_I \Delta_I}{\nu_{\text{tot}}} = \kappa_\pm \Delta_R + (1 - \kappa_\pm) \Delta_I.$$

First, assume that the attention weights are so that $\rho = \rho_y \ge 0$. In this case (and for this particular label y),

- When $\delta = 0$, we choose the relevance set $\mathcal{R}_+$ which ensures $\kappa_+ \ge \zeta \ge 0$ and $f_0 > 0$.
- When $\delta = \Delta > \Delta_0$, we choose the relevance set $\mathcal{R}_-$ which ensures $\kappa_- \le \zeta$ and $f_\Delta < 0$.

Secondly, assume that the attention weights are so that $\rho = \rho_y \le 0$. In this case,

- When $\delta = 0$, we choose the relevance set $\mathcal{R}_-$ which ensures $\kappa_- \le \zeta \le 1$ and $f_0 > 0$.
- When $\delta = \Delta > \Delta_0$, we choose the relevance set $\mathcal{R}_+$ which ensures $\kappa_+ \ge \zeta$ and $f_\Delta < 0$.

In either case, by adaptively choosing $\mathcal{R} \in \{\mathcal{R}_+, \mathcal{R}_-\}$ as a function of (δ, y) pair, we ensure accuracy is at most 50% because f_Δ and f_0 have conflicting signs. □

H.3 Success Proof for $\mathcal{R}$-Adaptive Self-Attention

Consider the setting of Theorem 1 and Appendix H.2. We have the following lemma which shows that self-attention can succeed in Theorem 1 if $\boldsymbol{U}$ can adapt to the relevance set (rather than $\mathcal{R}$ being adversarial to $\boldsymbol{U}$).

Lemma 15. *In (DATA), choose (δ^q, δ^w) to be $(0,0)$ or (Δ, Δ) equally-likely. Consider the self-attention model $f^{\text{SATT}}(\boldsymbol{U}, \boldsymbol{W})$ where we set*

$$\boldsymbol{U} = \mathbb{1}_{\mathcal{R}} \boldsymbol{w}_\star'^{\top} \quad \text{and} \quad \boldsymbol{W} = \Gamma \boldsymbol{I}.$$

This model achieves perfect accuracy whenever $\boldsymbol{w}_\star' = (\boldsymbol{I} - \bar{\boldsymbol{q}}_\star \bar{\boldsymbol{q}}_\star^\top)\boldsymbol{w}_\star \neq 0$ *by choosing*

$$\Gamma > \frac{1}{(1+\Delta)(\|\boldsymbol{q}_\star\| - \|\boldsymbol{w}_\star\|)^2 + \|\boldsymbol{w}_\star\|\|\boldsymbol{q}_\star\|(1 - |\rho(\boldsymbol{q}_\star, \boldsymbol{w}_\star)|)} \log(\Delta \frac{1-\zeta}{\zeta}).$$

where $\rho(\cdot)$ *is the correlation coefficient.*[3]

Proof. Thanks to the masking $\mathbb{1}_{\mathcal{R}}$, we only need to consider the attention scores along relevant tokens. Let $c = \|y\boldsymbol{w}_\star + \boldsymbol{q}_\star\|^2$. For each relevant token, the attention rows are given by

$$\boldsymbol{a}_i = \begin{cases} e^{\Gamma c} & \text{if} \quad i \in \mathcal{R} \\ e^{-\Delta\Gamma c} & \text{if} \quad i \notin \mathcal{R}. \end{cases}$$

To proceed, attention tokens corresponding to relevant tokens are given by

$$\boldsymbol{f} = \sum_{i \in \mathcal{R}} \boldsymbol{a}_i(\boldsymbol{w}_\star + y\boldsymbol{q}_\star) - \sum_{i \notin \mathcal{R}} \Delta \boldsymbol{a}_i(y\boldsymbol{w}_\star + \boldsymbol{q}_\star) \tag{88}$$

$$= (\zeta e^{\Gamma c} - \Delta(1-\zeta)e^{-\Delta\Gamma c})(y\boldsymbol{w}_\star + \boldsymbol{q}_\star). \tag{89}$$

Thus, using $\boldsymbol{w}_\star' \boldsymbol{w}_\star > 0$,

$$\texttt{sign}(y f^{\text{SATT}}(\boldsymbol{U}, \boldsymbol{W})) = \texttt{sign}(y\boldsymbol{w}_\star'^\top \boldsymbol{f}) = \texttt{sign}(\zeta e^{\Gamma c} - \Delta(1-\zeta)e^{-\Delta\Gamma c}).$$

Thus, we need $e^{(1+\Delta)\Gamma c} > \Delta\frac{1-\zeta}{\zeta}$ which is implied by $\Gamma > \frac{1}{(1+\Delta)c}\log(\Delta\frac{1-\zeta}{\zeta})$. To conclude, note that for both $y = \pm 1$

$$c \geq \|y\boldsymbol{w}_\star + \boldsymbol{q}_\star\|^2 \geq \|\boldsymbol{q}_\star\|^2 + \|\boldsymbol{w}_\star\|^2 - 2|\boldsymbol{q}_\star^\top \boldsymbol{w}_\star| \geq (\|\boldsymbol{q}_\star\| - \|\boldsymbol{w}_\star\|)^2 + \|\boldsymbol{w}_\star\|\|\boldsymbol{q}_\star\|(1 - |\rho(\boldsymbol{q}_\star, \boldsymbol{w}_\star)|) > 0.$$

where we used $|\boldsymbol{q}_\star^\top \boldsymbol{w}_\star| = \|\boldsymbol{q}_\star\|\|\boldsymbol{w}_\star\||\rho(\boldsymbol{q}_\star, \boldsymbol{w}_\star)|$. □

I Proofs for Sharp Characterization of Population Risk (Theorem 4)

Throughout this section, we use slightly different notation from the one stated in the main body for compactness purposes. Specifically, we set $Q = \|\boldsymbol{q}_\star\|^2, W = T\|\boldsymbol{w}_\star\|^2$ rather than $Q = \|\boldsymbol{q}_\star\|, W = \|\boldsymbol{w}_\star\|$.

Theorem 9. *Consider the prompt-attention model* $f_{\boldsymbol{\theta}}^{\text{ATT}}$. *Set* $Q = \|\boldsymbol{q}_\star\|^2, W = T\|\boldsymbol{w}_\star\|^2$, *suppose* $\boldsymbol{w}_\star \perp \boldsymbol{q}_\star$, *and let* $\tau, \bar{\tau} > 0$ *be hyperparameters. Consider the following algorithm which uses the hindsight knowledge of* $\boldsymbol{q}_\star$ *to estimate* $\boldsymbol{w}_\star$ *and make prediction:*

1. $\hat{\boldsymbol{w}} = (\boldsymbol{I} - \bar{\boldsymbol{q}}_\star \bar{\boldsymbol{q}}_\star)\nabla \mathcal{L}_{\boldsymbol{w}}(0, \tau\bar{\boldsymbol{q}}_\star)$.
2. *Set* $\boldsymbol{\theta} = (\hat{\boldsymbol{w}}, \bar{\tau}\bar{\boldsymbol{q}}_\star)$.

Suppose $\zeta^2 W, 1-\zeta, \alpha := n/d, e^Q, e^\tau$ *each lie between two positive absolute constants. Suppose* T *is polynomially large in* n *and these constants and* $\tilde{\mathcal{O}}(\cdot)$ *hides polynomial terms in* n. *Define* inverse-signal-to-noise-ratio: $\mathit{ISNR}(\alpha, \tau) = \frac{(1-\zeta)e^{2\tau(\tau - \sqrt{Q})}}{\zeta^2 W \alpha}$. *With probability* $1 - 2e^{-t^2/2} - \tilde{\mathcal{O}}(T^{-1/3})$ *over the training data, the test error obeys*

$$\text{ERR}(f_{\boldsymbol{\theta}}^{\text{ATT}}) = \mathcal{Q}\left(\frac{e^{\sqrt{Q}\bar{\tau} - \bar{\tau}^2}}{\sqrt{1 + (1 \mp \frac{1+t}{\sqrt{d}})\mathit{ISNR}(\alpha, \tau)}} \cdot \sqrt{\frac{\zeta^2 W}{1-\zeta}}\right) \pm \tilde{\mathcal{O}}(T^{-1/3}).$$

Above, $\mp, \pm$ *highlights the upper/lower range of the test error (see* (93) *for exact statement). In the limit* $T, d \to \infty$, *the test error converges in probability to*

$$\text{ERR}(\alpha, \zeta, Q, W, \tau, \bar{\tau}) = \mathcal{Q}\left(\frac{e^{\sqrt{Q}\bar{\tau} - \bar{\tau}^2}}{\sqrt{1 + \mathit{ISNR}(\alpha, \tau)}} \cdot \sqrt{\frac{\zeta^2 W}{1-\zeta}}\right)$$

[3]Note that the only instance Γ does not exist is when $\boldsymbol{q}_\star = c\boldsymbol{w}_\star$ for $|c| \geq 1$. In this scenario, classification is impossible using the linear head $\boldsymbol{w}_\star'$ without a bias term because all tokens are in the $\texttt{sign}(c)$ direction regardless of the label y.

In this limit, optimal hyperparameters are $\tau = \bar{\tau} = \sqrt{Q}/2$ *and leads to optimal* $\texttt{ISNR}(\alpha) := \frac{(1-\zeta)e^{-Q/2}}{\zeta^2 W \alpha}$ *and the error*

$$\textsc{err}(\alpha, \zeta, Q, W) = \mathcal{Q}\left(\frac{e^{Q/4}}{\sqrt{1+\texttt{ISNR}(\alpha)}} \cdot \sqrt{\frac{\zeta^2 W}{1-\zeta}}\right)$$

Proof. Without losing generality, assume first ζT tokens are relevant and remaining tokens are irrelevant. Consider $\boldsymbol{X}_I$ of size $(1-\zeta)T \times d$ induced by the irrelevant tokens with normal distribution. Observe that $\boldsymbol{g} = \boldsymbol{X}_I \bar{\boldsymbol{w}}_\star$ and $\mathbf{h} = \boldsymbol{X}_I \bar{\boldsymbol{q}}_\star$ are two independent i.i.d. $\mathcal{N}(0, \boldsymbol{I}_{(1-\zeta)T})$ vectors. Also for standard normal $g \sim \mathcal{N}(0,1)$, recall that moment-generating function is given by $\mathbb{E}[e^{\tau g}] = e^{\tau^2/2}$.

Step 1: Characterizing the distribution of $\hat{\boldsymbol{w}}$. Note that, the attention weights have the form $\boldsymbol{a} = \phi(\tau \begin{bmatrix} \sqrt{Q}\mathbb{1}_{\zeta T} \\ \mathbf{h} \end{bmatrix})$. Here, the softmax denominator is $T \cdot D_T$ where $D_T := (\zeta e^{\sqrt{Q}\tau} + \frac{1}{T}\sum_{i=1}^{(1-\zeta)T} e^{\tau h_i})$. Define $e^{\tau \mathbf{h}}$ to be the numerator corresponding to irrelevant tokens i.e.

$$e^{\tau \mathbf{h}} = [e^{\tau h_1} \ \ldots \ e^{\tau h_{(1-\zeta)T}}].$$

Define the matrix $\boldsymbol{Q}_\perp = \boldsymbol{I} - \bar{\boldsymbol{q}}_\star \bar{\boldsymbol{q}}_\star^\top$, $\boldsymbol{W}_\perp = \boldsymbol{I} - \bar{\boldsymbol{w}}_\star \bar{\boldsymbol{w}}_\star^\top$. Set the vector $\boldsymbol{v} = \frac{1}{T}\boldsymbol{Q}_\perp \boldsymbol{X}_I^\top e^{\tau \mathbf{h}}$ and $\boldsymbol{v}_\perp = \frac{1}{T}\mathbf{h}^\top e^{\tau \mathbf{h}} \bar{\boldsymbol{q}}_\star$. To proceed, observe that, for a single sample $(y, \boldsymbol{X})$, the gradient has the form

$$\nabla \mathcal{L}_{\boldsymbol{w}}^{y,\boldsymbol{X}}(0, \tau \bar{\boldsymbol{q}}_\star) = y \boldsymbol{X}^\top \boldsymbol{a} = \frac{\zeta(\boldsymbol{w}_\star + y\boldsymbol{q}_\star)e^{\sqrt{Q}\tau} + \boldsymbol{v} + \boldsymbol{v}_\perp}{D_T}. \tag{90}$$

After projection this onto the $\boldsymbol{q}_\star$-complement $\boldsymbol{Q}_\perp$, we get rid of the $\boldsymbol{q}_\star$ direction to obtain

$$\hat{\boldsymbol{w}}_{y,\boldsymbol{X}} = \boldsymbol{Q}_\perp \nabla \mathcal{L}_{\boldsymbol{w}}^{y,\boldsymbol{X}}(0, \tau \bar{\boldsymbol{q}}_\star) = \frac{1}{D_T}[\zeta \boldsymbol{w}_\star e^{\sqrt{Q}\tau} + \boldsymbol{Q}_\perp \boldsymbol{X}_I^\top e^{\tau \mathbf{h}}/T].$$

The projected gradient over the full dataset is given by the empirical average

$$\hat{\boldsymbol{w}} = \boldsymbol{Q}_\perp \nabla \mathcal{L}_{\boldsymbol{w}}(0, \tau \bar{\boldsymbol{q}}_\star) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{D_{i,T}}[\zeta \boldsymbol{w}_\star e^{\sqrt{Q}\tau} + \boldsymbol{Q}_\perp \boldsymbol{X}_{i,I}^\top e^{\tau \mathbf{h}_i}/T].$$

Here $\mathbf{h}_i, \boldsymbol{X}_{i,I}, D_{i,T}$ denote the random variables induced by the ith sample. Here, a critical observation is the fact that $\boldsymbol{Q}_\perp \boldsymbol{X}_{i,I}$ is independent of $\mathbf{h}_i$ (thanks to Gaussian orthogonality), thus, $\boldsymbol{Q}_\perp \boldsymbol{X}_{i,I} e^{\tau \mathbf{h}_i}$ is normal conditioned on $\mathbf{h}_i$. To proceed, we apply Chebyshev's inequality over number of tokens T. Recall that we assumed $e^\tau \le C$ for an absolute constant $C \ge 1$. This means that $e^{c\tau^2} \le C^{c\tau} \le C^{c \log C}$ is polynomial in C and is also upper bounded by a constant. In what follows $\tilde{\mathcal{O}}(\cdot)$ only reflects the T dependence and subsumes polynomial dependence on the terms n, C. For all $1 \le i \le n$, applying Chebyshev's inequality, for $T \gtrsim \text{poly}(n, e^{\tau^2})$, with probability $1 - T^{-1/3}$, we have that

- Since $\|e^{\tau \mathbf{h}_i}\|^2/T = \frac{1}{T}\sum_{j=1}^{T} e^{2\tau \mathbf{h}_{ij}}$ thus $\|e^{\tau \mathbf{h}_i}\|^2/T - (1-\zeta)e^{2\tau^2} \le \tilde{\mathcal{O}}(T^{-1/3})$,
- Set $\mathbb{E}[D_T] = D_\infty := \zeta e^{\sqrt{Q}\tau} + (1-\zeta)e^{\tau^2/2}$. $|D_{i,T} - D_\infty| \le \tilde{\mathcal{O}}(T^{-1/3})$.

With these, set $\boldsymbol{b}_i = \frac{\sqrt{1-\zeta}e^{\tau^2}}{\|e^{\tau \mathbf{h}_i}\|} e^{\tau \mathbf{h}_i}$ which is a vector with fixed ℓ_2 norm that is perfectly parallel to $e^{\tau \mathbf{h}_i}$. Since $\|\boldsymbol{b}_i\|^2 = \mathbb{E}[\|e^{\tau \mathbf{h}_i}\|^2/T] = (1-\zeta)e^{2\tau^2}$, from above, observe that,

$$\|\boldsymbol{b}_i - \frac{1}{\sqrt{T}} e^{\tau \mathbf{h}_i}\| \le \tilde{\mathcal{O}}(T^{-1/3}).$$

Now, let

$$\bar{\boldsymbol{v}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \boldsymbol{Q}_\perp \boldsymbol{X}_{i,I}^\top \boldsymbol{b}_i.$$

Since $\boldsymbol{Q}_\perp \boldsymbol{X}_{i,I}^\top, \boldsymbol{b}_i$ are independent and $\boldsymbol{b}_i$ has fixed ℓ_2 norm, we have that

$$\bar{\boldsymbol{v}} \sim \mathcal{N}(0, (1-\zeta)e^{2\tau^2} \boldsymbol{Q}_\perp).$$

Finally, let $\boldsymbol{c} = \zeta e^{\sqrt{Q}\tau}\boldsymbol{w}_\star$. Recalling $\sqrt{T}\|\boldsymbol{w}_\star\| = W$, combining the perturbations bounds above, we have that

$$\sqrt{T}\|\boldsymbol{c}/D_\infty - \frac{1}{n}\sum_{i=1}^{n}\frac{1}{D_{i,T}}(\zeta \boldsymbol{w}_\star e^{\sqrt{Q}\tau})\| \le \tilde{\mathcal{O}}(T^{-1/3}).$$

Combining these observe that

$$\|\sqrt{T}D_\infty \hat{\boldsymbol{w}} - \sqrt{T}\zeta e^{\sqrt{Q}\tau}\boldsymbol{w}_\star - \bar{\boldsymbol{v}}/\sqrt{n}\| \le \tilde{\mathcal{O}}(T^{-1/3}). \tag{91}$$

Since $\bar{\boldsymbol{v}}$ is normally distributed, above also implies that $\sqrt{T}D_\infty\hat{\boldsymbol{w}}$ converges to the normal distribution $\mathcal{N}(\sqrt{T}\zeta e^{\sqrt{Q}\tau}\boldsymbol{w}_\star, \frac{(1-\zeta)e^{2\tau^2}}{n}\boldsymbol{Q}_\perp)$ in the limit $T\to\infty$.

Lemma 16 (Inverse Signal-to-Noise Ratio (ISNR))**.** *Set $\boldsymbol{W}_\perp = \boldsymbol{I} - \bar{\boldsymbol{w}}_\star\bar{\boldsymbol{w}}_\star^\top$. Define SNR of $\hat{\boldsymbol{w}}$ to be*

$$\mathtt{ISNR}(\hat{\boldsymbol{w}}) = \frac{\|\boldsymbol{W}_\perp\hat{\boldsymbol{w}}\|^2}{\|\bar{\boldsymbol{w}}_\star^\top\hat{\boldsymbol{w}}\|^2}.$$

Recall $\mathtt{ISNR}(\alpha,\tau) = \frac{(1-\zeta)e^{2\tau(\tau-\sqrt{Q})}}{\zeta^2 W\alpha}$. With probability $1-2e^{-t^2/2}-T^{-1/3}$ over the dataset, we have that

$$\left(1-\frac{t+1}{\sqrt{d}}-\tilde{\mathcal{O}}(T^{-\frac{1}{3}})\right)_+^2 \le \frac{\mathtt{ISNR}(\hat{\boldsymbol{w}})}{\mathtt{ISNR}(\alpha,\tau)} \le \left(1+\frac{\tau}{\sqrt{d}}+\tilde{\mathcal{O}}(T^{-\frac{1}{3}})\right)^2.$$

Proof. Let us recall the standard normal concentration: For $\boldsymbol{g}\sim\mathcal{N}(0,\boldsymbol{I}_{d-1})$, $\sqrt{d-1}\ge\mathbb{E}[\|\boldsymbol{g}\|]\ge\frac{d-1}{\sqrt{d}}$. Thus, with probability $1-2e^{-t^2/2}$, through Lipschitz concentration,

$$\sqrt{d}+t\ge\|\boldsymbol{g}\|\ge\sqrt{d}-1-t.$$

This means that, with the same probability

$$\sqrt{d}+t\ge\frac{\|\bar{\boldsymbol{v}}\|}{\sqrt{1-\zeta e^{\tau^2}}}\ge(\sqrt{d}-1-t)_+.$$

We first upper bound $\|\boldsymbol{W}_\perp\hat{\boldsymbol{w}}\|^2$. Recalling (91),

$$\|\boldsymbol{W}_\perp\hat{\boldsymbol{w}}-\bar{\boldsymbol{v}}/\sqrt{n}\|\le\tilde{\mathcal{O}}(T^{-1/3}).$$

Thus,

$$(\sqrt{d}+t)^2+\tilde{\mathcal{O}}(T^{-1/3})\ge\frac{n\|\boldsymbol{W}_\perp\hat{\boldsymbol{w}}\|^2}{(1-\zeta)e^{2\tau^2}}\ge(\sqrt{d}-1-t)_+^2-\tilde{\mathcal{O}}(T^{-1/3}).$$

Using $\|\boldsymbol{w}_\star\|^2T=W$, We similarly have that

$$|\|\bar{\boldsymbol{w}}_\star^\top\hat{\boldsymbol{w}}\|^2-W\zeta^2e^{2\sqrt{Q}\tau}|\le\tilde{\mathcal{O}}(T^{-1/3}).$$

To conclude, with probability $1-2e^{-t}-T^{-1/3}$, $\mathtt{ISNR}(\hat{\boldsymbol{w}})$ obeys

$$\frac{(\sqrt{d}+t)^2+\tilde{\mathcal{O}}(T^{-1/3})}{W\zeta^2e^{2\sqrt{Q}\tau}-\tilde{\mathcal{O}}(T^{-1/3})}\ge\frac{n}{(1-\zeta)e^{2\tau^2}}\mathtt{ISNR}(\hat{\boldsymbol{w}})\ge\frac{(\sqrt{d}-1-t)_+^2-\tilde{\mathcal{O}}(T^{-1/3})}{W\zeta^2e^{2\sqrt{Q}\tau}+\tilde{\mathcal{O}}(T^{-1/3})}$$

Rewriting this bound, we find

$$\left(1+\frac{t}{\sqrt{d}}+\tilde{\mathcal{O}}(T^{-\frac{1}{3}})\right)^2\frac{(1-\zeta)e^{2\tau^2}}{\zeta^2W\alpha e^{2\sqrt{Q}\tau}}\ge\mathtt{ISNR}(\hat{\boldsymbol{w}})\ge\left(1-\frac{1+t}{\sqrt{d}}-\tilde{\mathcal{O}}(T^{-\frac{1}{3}})\right)_+^2\frac{(1-\zeta)e^{2\tau^2}}{\zeta^2W\alpha e^{2\sqrt{Q}\tau}}.$$

Recalling the definition of $\mathtt{ISNR}(\alpha,\tau)=\frac{(1-\zeta)e^{2\tau(\tau-\sqrt{Q})}}{\zeta^2W\alpha}$, we conclude with the bound. □

Step 2: Characterizing the error rate of $\boldsymbol{\theta} = (\hat{\boldsymbol{w}}, \bar{\tau}\boldsymbol{q}_\star)$. To achieve this goal, we will leverage Theorem 10. Since conditions of this theorem is satisfied (noticing that their γ is our $\mathtt{ISNR}(\hat{\boldsymbol{w}})$ which is upper bounded by a positive constant), for a new test point $(y,\boldsymbol{X})$, we have that

$$\left|\mathrm{ERR}(f_{\boldsymbol{\theta}}^{\mathrm{ATT}})-\mathcal{Q}\left(\frac{e^{\sqrt{Q}\bar{\tau}-\bar{\tau}^2}}{\sqrt{1+\mathtt{ISNR}(\hat{\boldsymbol{w}})}}\cdot\sqrt{\frac{\zeta^2W}{1-\zeta}}\right)\right|\le\tilde{\mathcal{O}}(T^{-1/3}).$$

Using the Lipschitzness of the Q-function (i.e. $\mathcal{Q}(x+\epsilon)-\mathcal{Q}(x)=\int_x^{x+\epsilon} e^{-t^2/2}dt \le \epsilon$), as we have done in Theorem 10, we pull out the perturbation term $\tilde{\mathcal{O}}(T^{-1/3})$ within $\texttt{ISNR}(\hat{\boldsymbol{w}})$ to obtain the advertised bound

$$\mathcal{Q}\left(\frac{e^{\sqrt{Q}\bar{\tau}-\bar{\tau}^2}}{\sqrt{1+(1+\frac{t}{\sqrt{d}})\texttt{ISNR}(\alpha,\tau)}}\cdot\sqrt{\frac{\zeta^2 W}{1-\zeta}}\right)-\tilde{\mathcal{O}}(T^{-1/3}) \le \texttt{ERR}(f_{\boldsymbol{\theta}}^{\texttt{ATT}}) \le \tag{92}$$

$$\mathcal{Q}\left(\frac{e^{\sqrt{Q}\bar{\tau}-\bar{\tau}^2}}{\sqrt{1+(1-\frac{1+t}{\sqrt{d}})_+\texttt{ISNR}(\alpha,\tau)}}\cdot\sqrt{\frac{\zeta^2 W}{1-\zeta}}\right)+\tilde{\mathcal{O}}(T^{-1/3}). \tag{93}$$

To emphasize, this bound holds with probability $1-2e^{-t^2/2}-\tilde{\mathcal{O}}(T^{-1/3})$ over a new test datapoint $(y,\boldsymbol{X})$. To see the optimal choices for $\bar{\tau},\tau$, we need to optimize the error bound. This results in

$$\bar{\tau}_\star = \arg\min_{\bar{\tau}} \sqrt{Q}\bar{\tau}-\bar{\tau}^2 = \sqrt{Q}/2 \tag{94}$$

$$\tau_\star = \arg\min_{\tau} \texttt{ISNR}(\alpha,\tau) = 2\tau(\tau-\sqrt{Q}) = \sqrt{Q}/2. \tag{95}$$

□

Theorem 10. *Consider the prompt-attention model $f_{\boldsymbol{\theta}}^{\texttt{ATT}}$ where we set $\boldsymbol{\theta}=(\boldsymbol{w}_\star+\boldsymbol{p},\tau\bar{\boldsymbol{q}}_\star)$. Set $Q=\|\boldsymbol{q}_\star\|^2, W=T\|\boldsymbol{w}_\star\|^2$. Here τ is a tuning parameter and $\boldsymbol{p}$ is a perturbation vector and assume all vectors are perpendicular i.e. $\boldsymbol{p}\perp\boldsymbol{w}_\star\perp\boldsymbol{q}_\star$. Set $\gamma:=\|\boldsymbol{p}\|^2/\|\boldsymbol{w}_\star\|^2$ and suppose $1+\gamma,\zeta^2 W,1-\zeta,e^Q,e^\tau$ each lie between two positive absolute constants. $\tilde{\mathcal{O}}(\cdot)$ subsumes polynomial dependencies in these constants. We have that*

$$\left|\texttt{ERR}(f_{\boldsymbol{\theta}}^{\texttt{ATT}})-\mathcal{Q}\left(\frac{e^{\sqrt{Q}\tau-\tau^2}}{\sqrt{1+\gamma}}\cdot\sqrt{\frac{\zeta^2 W}{1-\zeta}}\right)\right| \le \mathcal{O}(T^{-1/3}).$$

Thus, as $T\to\infty$, the optimal tuning obeys $\tau_\star=\sqrt{Q}/2$ and yields an error of $\mathcal{Q}\left(e^{Q/4}\cdot\sqrt{\frac{\zeta^2 W}{1-\zeta}}\right)$.

Proof. Let us recap the notation of Theorem 4. Without losing generality, assume first ζT tokens are relevant and remaining tokens are irrelevant. Consider $\boldsymbol{X}_I$ of size $(1-\zeta)T\times d$ induced by the irrelevant tokens with normal distribution. Using orthogonality of $\boldsymbol{q}_\star,\boldsymbol{w}_\star,\boldsymbol{p}$, observe that $\boldsymbol{g}=\boldsymbol{X}_I(\bar{\boldsymbol{w}}_\star+\frac{\boldsymbol{p}}{\|\boldsymbol{w}_\star\|}) \sim \mathcal{N}(0,(1+\gamma)\boldsymbol{I}_{(1-\zeta)T})$ and $\mathbf{h}=\boldsymbol{X}_I\bar{\boldsymbol{q}}_\star\sim\mathcal{N}(0,\boldsymbol{I}_{(1-\zeta)T})$ are independent vectors. Also for standard normal $g\sim\mathcal{N}(0,1)$, recall that moment-generating function is given by $\mathbb{E}[e^{\tau g}]=e^{\tau^2/2}$.

Note that, the attention weights have the form $\boldsymbol{a}=\phi(\tau\begin{bmatrix}\sqrt{Q}\mathbb{1}_{\zeta T}\\ \mathbf{h}\end{bmatrix})$. Here, the softmax denominator is $T\cdot D_T$ where $D_T:=(\zeta e^{\sqrt{Q}\tau}+\frac{1}{T}\sum_{i=1}^{(1-\zeta)T}e^{\tau h_i})$. Define $e^{\tau\mathbf{h}}$ to be the numerator corresponding to irrelevant tokens i.e.

$$e^{\tau\mathbf{h}}=\begin{bmatrix}e^{\tau h_1} & \dots & e^{\tau h_{(1-\zeta)T}}\end{bmatrix}.$$

Define the matrix $\boldsymbol{Q}_\perp=\boldsymbol{I}-\bar{\boldsymbol{q}}_\star\bar{\boldsymbol{q}}_\star^\top$, $\boldsymbol{W}_\perp=\boldsymbol{I}-\bar{\boldsymbol{w}}_\star\bar{\boldsymbol{w}}_\star^\top$. To proceed, observe that, the prediction with $\boldsymbol{\theta}=(\boldsymbol{w}_\star+\boldsymbol{p},\tau\bar{\boldsymbol{q}}_\star)$ is given by

$$\frac{D_T}{\sqrt{T}\|\boldsymbol{w}_\star\|}yf_{\boldsymbol{\theta}}^{\texttt{ATT}}(\boldsymbol{X})=\sqrt{T}(\bar{\boldsymbol{w}}_\star+\frac{\boldsymbol{p}}{\|\boldsymbol{w}_\star\|})^\top[\zeta e^{\|\boldsymbol{q}_\star\|\tau}(\boldsymbol{w}_\star+y\boldsymbol{q}_\star)+\frac{\boldsymbol{X}_I e^{\tau\mathbf{h}}}{T}] \tag{96}$$

$$=\zeta e^{\|\boldsymbol{q}_\star\|\tau}\sqrt{W}+\frac{1}{\sqrt{T}}\boldsymbol{g}^\top e^{\tau\mathbf{h}}. \tag{97}$$

With this, conditioned on $e^{\tau\mathbf{h}}$ observe that $\boldsymbol{g}^\top e^{\tau\mathbf{h}}\sim\mathcal{N}(0,\frac{1}{T}\|e^{\tau\mathbf{h}}\|^2)$, thus,

$$\mathbb{P}_{\boldsymbol{g}}(yf_{\boldsymbol{\theta}}^{\texttt{ATT}}(\boldsymbol{X})>0)=1-\mathcal{Q}\left(\frac{\zeta e^{\sqrt{Q}\tau}\sqrt{W}}{\sqrt{1+\gamma}\|e^{\tau\mathbf{h}}\|/\sqrt{T}}\right).$$

To proceed, similar to Theorem 4, we apply Chebyshev's inequality over number of tokens T to find that with probability $1-\tilde{\mathcal{O}}(T^{-1/3})$ over $\mathbf{h}$,

$$|\|e^{\tau\mathbf{h}}\|^2/T-e^{2\tau^2}|\le\tilde{\mathcal{O}}(T^{-1/3}).$$

In aggregate, this implies that, with probability $1 - \tilde{\mathcal{O}}(T^{-1/3})$ over $\mathbf{h}$, we have that

$$1-\mathcal{Q}\left((1+\tilde{\mathcal{O}}(T^{-1/3}))\frac{\zeta e^{\sqrt{Q}\tau}\sqrt{W}}{\sqrt{1+\gamma}e^{\tau^2}}\right) \geq \mathbb{P}_{\boldsymbol{g}}(y f_{\boldsymbol{\theta}}^{\text{ATT}}(\boldsymbol{X}) > 0) \geq 1-\mathcal{Q}\left((1-\tilde{\mathcal{O}}(T^{-1/3}))\frac{\zeta e^{\sqrt{Q}\tau}\sqrt{W}}{\sqrt{1+\gamma}e^{\tau^2}}\right),$$

Finally, note that since $\frac{\zeta e^{\sqrt{Q}\tau}\sqrt{W}}{\sqrt{1+\gamma}e^{\tau^2}}$ is upper/lower bounded by a positive constant, and since $\mathcal{Q}(x+\epsilon) - \mathcal{Q}(x) = \int_x^{x+\epsilon} e^{-t^2/2}dt \leq \epsilon$, we can rewrite

$$\left|\mathbb{P}_{\boldsymbol{g}}(y f_{\boldsymbol{\theta}}^{\text{ATT}}(\boldsymbol{X}) > 0) - \mathcal{Q}\left(\frac{\zeta e^{\sqrt{Q}\tau}\sqrt{W}}{\sqrt{1+\gamma}e^{\tau^2}}\right)\right| \leq \tilde{\mathcal{O}}(T^{-1/3}).$$

Union bounding with failure probability over $\mathbf{h}$, we conclude with the result. □

J USEFUL FACTS

For a random variable Z and $\alpha > 0$, $\|Z\|_{\psi_\alpha}$ denotes its ψ_α-norm for Orlicz function $\psi_\alpha(z) = e^{z^\alpha} - 1$ Ledoux and Talagrand (1991).

Fact J.1. *Let $X_1, \ldots, X_n$ be independent zero-mean sub-gaussian or sub-exponential random variables with $\|X_i\|_{\psi_m} \leq K$ for all $i \in [n]$ for either $m = 2$ or $m = 1$. Then,*

$$\left\|\frac{1}{n}\sum_{i\in[n]} X_i\right\|_{\psi_m} \leq \frac{CK}{\sqrt{n}}.$$

Fact J.2. *Pollard (1990) The following identity holds for Orlitz norms*

$$\|XY\|_{\psi_{\frac{\alpha\beta}{\alpha+\beta}}} \leq c\|X\|_{\psi_\alpha} \cdot \|Y\|_{\psi_\beta} \tag{98}$$

for a fixed numerical constant c.

Next we state a Lemma from Talagrand quoted directly from Lemma 22 of Mohammadi et al. (2019).

Fact J.3. *Ledoux and Talagrand (1991)For any scalar $\alpha \in (0, 1]$, there exists a constant C_α such that for any sequence of independent random variables $\xi_1, \xi_2, \ldots, \xi_N$ we have*

$$\left\|\sum_i \xi_i - \mathbb{E}\left[\sum_i \xi_i\right]\right\|_{\psi_\alpha} \leq C_\alpha \left(\max_i \|\xi_i\|_{\psi_\alpha}\right)\sqrt{N}\log N.$$

Fact J.4. *Let $\boldsymbol{z} \in \mathbb{R}^d$ be zero-mean k-subgausssian vector. Then, the following are true*

$$\mathbb{P}\left(\|\boldsymbol{z}\| \geq t\right) \leq 2e^{-t^2/(16dK^2)}$$